Machine Learning and Best Practices for Data Analysis

A Comparison of Commonly Used Models

Henry Santangelo

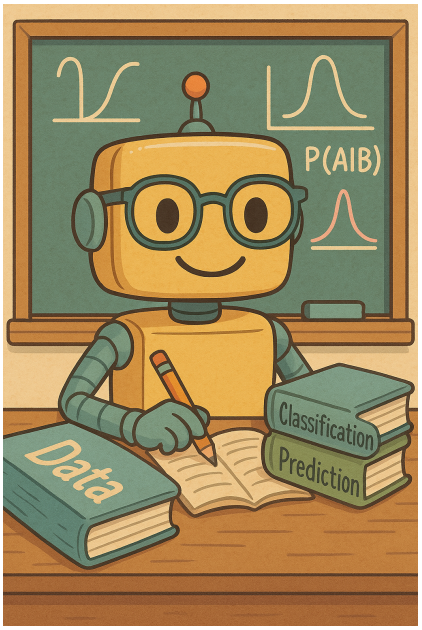

Image generated using ChatGPT by OpenAI

Mathematics
Greenwich Country Day School
4/14/2025

Contents

1 Introduction

Machine learning is "a subfield of artificial intelligence that gives computers the ability to learn without explicitly being programmed"[1]. Computers have the ability to learn through developed models. Typically, the "learning" of a model is tuning weights and biases, which are just numbers that dictate a model's output. The learning of a model happens through minimizing some error function by the tuning of weights and biases. The error function defines a model's behavior. If the function is flawed, then so will the model. For example if a self driving car only has error when it hits an object, then it might learn to never move. Within machine learning there are two overarching branches: classification and prediction.

Classification seeks "to identify mathematical and/or statistical relationships between independent variables (discrete or continuous) that can effectively distinguish or characterize various levels with a nominal dependent variable (categorical variable)"[2]. These independent variables are used as inputs to a model and the categorical/dependent variable is the output. For example, measurements of a tumor could be the independent variables (weight, height, width, etc.) and the categorical variable could whether the tumor is cancerous (malignant or benign).

Prediction seeks to predict a dependent continuous variable (result) from independent variables. For example, the independent variables could be measurements of a home (square footage, land area, number of bedrooms) and the resulting variable could be the price of a home.

As the use of machine learning becomes increasingly more popular for data analysis, it is important that researchers and data analysts follow the best practices of analysis. Choosing what model is right for the job is equally important. This paper seeks to uncover what models should be used in a

1. Sara Brown. 2021. *Machine learning, explained.* MIT Management Sloan School. Last modified April 2021. Accessed December 10, 2024. This source explains machine learning concepts, applications, and types. Published by MIT, it provides a strong foundation for understanding the field. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

2. Brian Carnahan, Gerard Meyer, and Lois-Ann Kuntz. 2003. Comparing statistical and machine learning classifiers: alternatives for predictive modeling in human factors research. This peer-reviewed source compares traditional statistical classifiers with machine learning approaches, such as decision trees and genetic programming, in human performance applications. *Human Factors* 45 (3): 408+.

given scenario, the weaknesses of models, and the strengths of models.

Overfitting is one weakness of some commonly used machine learning models. Overfitting is the process of which a model learns noise of the training data instead of just the patterns. This behavior makes a model less reliable[3].

1.1 Classification Models

Commonly used classification models include logistic regression, discriminant analysis, and naive Bayesian classification. These models are all easily accessible in almost every programming language and data analysis software.

Logistic regression uses the sigmoid function to plot data points. For simplicity, logistic regression will only be used with two categories, a positive category (the second category) and a negative category (the first category), e.g. malignant and benign. The sigmoid function looks like an "s" curve.

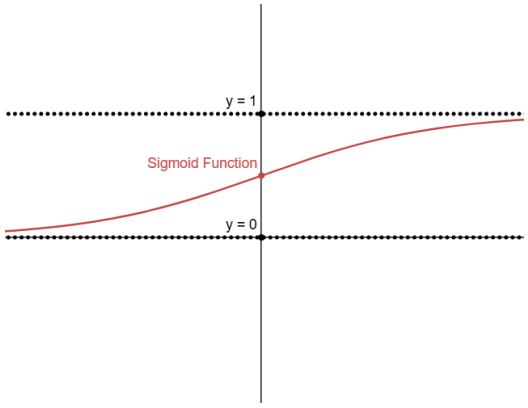

Figure 1: Sigmoid Function

A linear combination of the independent variables is used as input to the sigmoid function to generate a confidence value. The confidence value measures how confident that the model is that a data point in a category. A value of 0.8 would mean the model predicts with 80% confidence that the data point is categorized into the second category. To continue with the example, if the confidence value was 0.8 , then the model would predict that the tumor

3. Amazon. *What is overfitting?* Accessed April 1, 2025. https://aws.amazon.com/what-is/overfitting/.

was malignant. Whereas a value of 0.2 would mean the model predicts with 20% confidence that the data point is categorized into the second category or an 80% confidence the data point is categorized into the first category. To continue with the example, if the confidence value was 0.2 , then the model would predict that the tumor was benign.

Logistic regression is the most widely used classification model. This popularity comes from the ease of use, it is built into many programming languages, and its ability to be general purpose. This model has been widely applied to fields like medicine, biology, physics, and marketing.

Discriminant Analysis uses linear planes, lines, and hyperplanes to separate data points into categories[4]. Originally developed by Sir Ronald Fisher, the technique has been popular in statistical analysis. Planes separate data points as shown in the figure below. This is an easy way to classify points but it does lack nuance due to the linear nature. The model works best when the groups of data points are most spread out from each other. It is one of the simplest classification models.

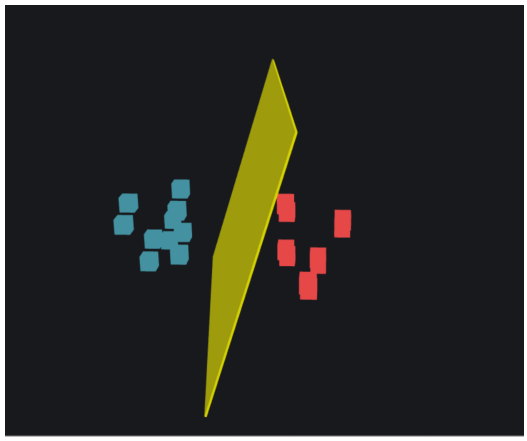

Figure 2: Linear Discriminant Analysis

4. IBM. 2023. *What is linear discriminant analysis (lda)?* https://www.ibm.com/think/topics/linear-discriminant-analysis. Last modified November 27, 2023. Accessed March 6, 2025. This source provides an informative introduction to LDA and its mathematical foundations.

The next classification model is a *Naive Bayes'* classifier. This classification model is based on conditional probability Bayes' rule. Bayes' rule states that

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Sometimes also written, in the more readable way,

$$\text{The Posterior Probability} = \frac{\text{Prior Probability} * \text{Likelihood}}{\text{Evidence}}$$

This theorem was developed by Thomas Bayes, an 18th century statistician and philosopher[5]. A Naive Bayes' classifier assumes that each independent variable is equally important which can be a disadvantage in capturing nuance. A Gaussian Naive Bayes' Classifier, which is what is implemented later, assumes the continuous independent variables follow a Gaussian distribution (also known as a normal distribution). It is a more outdated model and can be derived from basic undergraduate statistics and does not require any heavy computation[6].

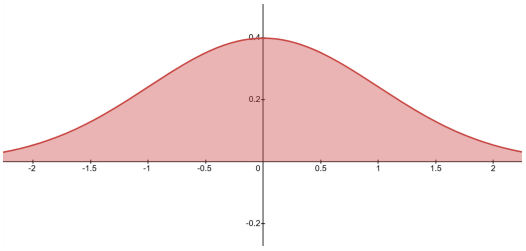

Figure 3: Gaussian (Normal) Distribution

5. D. R. Bellhouse. *The reverend thomas bayes, frs: a biography to celebrate the tercentenary of his birth.* Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University. Accessed March 6, 2025. This source provides information on Thomas Bayes and his contributions to mathematics and statistics, emphasizing the significance of Bayes' rule. https://biostat.jhsph.edu/courses/bio621/misc/bayesbiog.pdf.

6. Kilian Weinberger. 2018. *Bayes classifier and naive bayes,* July. Accessed April 1, 2025. https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote05.html; York Yong. *Introduction to naive bayes algorithm - gaussian and multinomial variants.* Accessed April 1, 2025. https://www.kaggle.com/discussions/general/468420.

1.2 Prediction Models

The prediction models that will be analyzed in this paper are *ordinary least squares regression*, *ridge regression*, and *LASSO regression*. They seek to find a linear model to predict a dependent variable using continuous independent variables.

When testing prediction models, it is important to break up the data into training and validation data. This is important so that the model is tested on different data than it is trained on. *Ordinary least squares regression* splits the data into training and validation sets of data. Ridge and LASSO regression split data into training, testing, and validation sets of data where the testing data is also used in the training but chooses a λ value instead of affecting β[7]. Each model, using various methods, is trained on the training data in order to minimize some error function. Then, the results of the training can be captured on the validation set, recording the R^2.

7. IBM. 2024. *What is lasso regression?,* January. Accessed April 1, 2025. https://www.ibm.com/think/topics/lasso-regression.

Part I
Technical Analysis

2 Least Squares Regression

This section explains the technical aspects behind ordinary least squares linear regression.

2.1 The Setup

Suppose we have data matrix

$$X = \left(\vec{x}_1, \ldots, \vec{x}_d, \vec{1}\right) \in \mathbb{R}^{n \times (d+1)},$$

where each $\vec{x}_i$ is a column vector representing one of the d predictor variables for all n observations, and the column $\vec{1}$ accounts for the intercept term in our model. The intercept term is b of $y = mx + b$. Also known as the bias term.

We also assume there exists a true set of regression coefficients given by

$$\vec{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{d+1} \end{pmatrix}.$$

These coefficients represent the influence each predictor has on the outcome. This is like m of $y = mx + b$.[8]

8. Cosma Shalizi. 2015. *Simple linear regression in matrix format.* Carnegie Mellon University Statistics and Data Science, Carnegie Mellon University. Last modified October 13, 2015. Accessed March 6, 2025. This lecture summary explains least squares regression using matrix notation. https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/13/lecture-13.pdf.

2.2 The Model

Our linear model assumes that the response variable $\vec{y}$ is described by

$$\vec{y} = X\vec{\beta} + \vec{\epsilon},$$

where $\vec{\epsilon}$ is an error term independent of X. This means that our true results $\vec{y}$ are a combination of the true linear relationship $X\vec{\beta}$ and some random error $\vec{\epsilon}$[9].

2.3 Finding the Best Fit

The goal of linear regression is to find an estimate $\vec{b}$ for $\vec{\beta}$ that makes our model's predictions as close as possible to the real data. We do this by minimizing the sum of the squared differences (error) between the true values $\vec{y}$ and the predicted values $X\vec{b}$. This sum is written as:

$$\|\vec{y} - X\vec{b}\|_2^2.$$

To achieve the smallest possible error, we project $\vec{y}$ onto the column space of X. This projection minimizes the length of $\vec{\epsilon}$ and ensures that it is orthogonal to every column in X.

2.4 Deriving the Coefficients

Assuming that the error vector is indeed orthogonal to the columns of X, we have:

$$X^T\vec{y} = X^TX\vec{b} + X^T\vec{\epsilon},$$

and since $X^T\vec{\epsilon} = 0$, this simplifies to:

$$X^T\vec{y} = X^TX\vec{b}.$$

To solve for $\vec{b}$, we multiply both sides by the inverse of X^TX, leading to the well known least squares solution:

$$\vec{b} = (X^TX)^{-1}X^T\vec{y}.$$

This equation gives us the best-fit coefficients for our linear model based on the input[10].

9. Shalizi 2015.
10. *Id.*

3 LASSO and Ridge Regression

This section explains why least squares may not be the best choice and some alternatives for linear regression.

3.1 Weakness of the ordinary least squares regression

Overfitting. To see how overfitting is in an issue in the solution to ordinary least squares, let's express X as its SVD (Singular Value Decomposition)

$$X = U^T \Sigma V,$$

where U, V are orthogonal matrices and Σ is the diagonal matrix of the singular values. Then consider

$$X^T = V^T \Sigma U$$

$$\Rightarrow X^T X = V^T \Sigma U U^T \Sigma V$$

$$\Rightarrow X^T X = V^T \Sigma \Sigma V$$

$$\Rightarrow X^T X = V^T \Sigma^2 V$$

$$\Rightarrow (X^T X)^{-1} = (V^T \Sigma^2 V)^{-1}$$

$$\Rightarrow (X^T X)^{-1} = V^T \Sigma^{2^{-1}} V$$

If the columns of X are almost linearly dependent, a singular value in X will be very small, which will cause numerical instability, because we are taking the inverse of the square of Σ. So this means that if the columns are fairly linearly dependent, it's possible that a small perturbation in the data will cause the smallest eigenvalue to change, which can cause a big change in the inverse. Hence we might be drastically changing $\vec{b}$ just because of noise.

This problem may seem unsolvable, however, there exist many other types of linear regression including ridge and LASSO regression. The ways the two mentioned models combat overfitting are different yet similar.

3.2 Ridge Regression

3.2.1 The Setup

We set up the same as ordinary least squares regression. With

$$X = \left(\vec{x}_1, \ldots, \vec{x}_d, \vec{1}\right) \in \mathbb{R}^{n\times(d+1)},$$

and

$$\vec{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{d+1} \end{pmatrix}.$$

3.2.2 The Model

Our linear model (same as least squares) assumes that the response variable $\vec{y}$ is described by

$$\vec{y} = X\vec{\beta} + \vec{\epsilon},$$

where $\vec{\epsilon}$ is an error term independent of X. This means that our true results $\vec{y}$ are a combination of the true linear relationship $X\vec{\beta}$ and some random error $\vec{\epsilon}$[11].

3.2.3 Finding the Best Fit

Instead of just minimizing the sum of the squared residuals, we minimize

$$g(\beta) = \lambda\beta^T\beta + \|y - X\beta\|^2$$

This extra term with λ helps prevent overfitting which is explained in the next section. λ is a constant which there exists an algorithm to choose it also described below.

3.2.4 Deriving the Coefficients

The derivation of these coefficients requires matrix calculus. The solution will be at the bottom of the section.

$$\nabla_\beta \, |y - X\beta\|^2 = -2X^T(y - X\beta)$$

11. Jacob Murel and Eda Kavlakoglu. 2023. *What is ridge regression?* Accessed November 21, 2023. https://www.ibm.com/think/topics/ridge-regression.

$$\nabla_\beta \ \lambda|\beta\|^2 = 2\lambda\beta$$

$$g(\beta) = \lambda\beta^T\beta + \|y - X\beta\|^2$$

$$\Rightarrow \nabla_\beta g(\beta) = 2\lambda\beta - 2X^T(y - X\beta)$$

To find a minimum using a gradient, you set it to 0 the same as a regular derivative[12].

$$0 = 2\lambda\beta - 2X^T(y - X\beta)$$

Divide out the 2 and move term over

$$X^T(y - X\beta) = \lambda\beta$$
$$X^Ty - X^TX\beta = \lambda\beta$$
$$X^Ty = \lambda\beta + X^TX\beta$$

Factor out β

$$X^Ty = (\lambda I + X^TX)\beta$$
$$(\lambda I + X^TX)^{-1}X^Ty = \beta$$

Switch equation sides

$$\beta = (\lambda I + X^TX)^{-1}X^Ty$$

This result for β is very similar to ordinary least squares regression, but there is λI term in the inverse. This fixes the previous problem by not allowing the values in the inverse to blow up. Each entry in the $\Sigma^2 + \lambda I$ matrix is at least λ which means the maximum eigenvalue of an entry in $(X^TX + \lambda I)^{-1}$ is $\frac{1}{\lambda}$. To see this more concretely, consider

$$X^TX + \lambda I = V^T\Sigma^2V + \lambda I$$
$$\Rightarrow X^TX + \lambda I = V^T\Sigma^2V + V^T\lambda IV$$
$$\Rightarrow X^TX + \lambda I = V^T(\Sigma^2 + \lambda I)V$$
$$\Rightarrow (X^TX + \lambda I)^{-1} = V^T(\Sigma^2 + \lambda I)^{-1}V.$$

In this case, $(X^TX+\lambda I)^{-1}$ won't "blow up" as in the normal linear regression case. This fixes the issue of having small eigenvalues when you have almost linearly dependent columns.

12. Murel and Kavlakoglu 2023.

3.2.5 Ridge Regression Algorithm

λ can also be optimized in a way. Consider the following way to pick λ. Split your data into 3 parts, training, testing, and validation. 60% of the data is training, 20% of the data is testing, and 20% of the data is validation. Use the training set data to learn the regression coefficients for different λ, e.g.,

$$\lambda = (0.001, 0.01, 0.1, \dots, 10000, 100000, \dots).$$

Then test each set of coefficients on the validation set, and then we keep the best one. Then we use this best λ on the testing set to report your results of the regression.

3.3 LASSO Regression

3.3.1 The Setup

We set up the same as ordinary least squares regression. With

$$X = \left(\vec{x}_1, \dots, \vec{x}_d, \vec{1}\right) \in \mathbb{R}^{n \times (d+1)},$$

and

$$\vec{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{d+1} \end{pmatrix}.$$

3.3.2 The Model

Our linear model (same as least squares and ridge) assumes that the response variable $\vec{y}$ is described by

$$\vec{y} = X\vec{\beta} + \vec{\epsilon},$$

where $\vec{\epsilon}$ is an error term independent of X. This means that our true results $\vec{y}$ are a combination of the true linear relationship $X\vec{\beta}$ and some random error $\vec{\epsilon}$.

3.3.3 Finding the Best Fit

Instead of just minimizing the sum of the squared residuals, we minimize

$$g(\beta) = \lambda\|\beta\|_1 + \|y - X\beta\|^2$$

This extra term with λ helps prevent overfitting which is explained in the next section. This is different from ridge regression as we use the L1 norm of β instead of the Squared Euclidean norm[13].

3.4 Deriving the coefficients

Unlike ridge regression, the L1 norm is not differentiable at points where $\beta = 0$. Therefore, we must use subgradient methods. The subgradient of $|\beta_j|$ with respect to β_j is

$$\frac{\partial}{\partial\beta_j} = \begin{cases} 1 & \beta_j > 0 \\ -1 & \beta_j < 0 \\ 0 & 0 = \beta_j \end{cases}$$

I have chosen 0 for when $\beta_j = 0$, however, any subgradient between -1 and 1 is valid.

Thus, the subgradient of $g(\beta)$ with respect to β is

$$\nabla_\beta g(\beta) = -2X^T(y - X\beta) + \lambda s$$

Where s is a vector with s_j chosen from the subgradient of β_j. Setting this subgradient to 0 does not yield a closed-form solution in general; however, there is a special case when the columns of X are orthonormal.

For this special case, the solution for each coefficient can be written using the soft-thresholding operator.

$$\hat{\beta}_j = S_\lambda(\hat{\beta}_j^{\,OLS})$$

Where

$$\hat{\beta}_j^{\,OLS}$$

is the ordinary least squares estimate for the jth component. And

$$S_\lambda(z) = \text{sign}(z)\max\{|z| - \lambda, 0\}$$

13. Andreas Tilevak. 2022. *Lasso regression - explained,* July. Accessed April 1, 2025. https://www.youtube.com/watch?v=bPFjfZWWQO0.

3.4.1 LASSO Regression Algorithm

Because the previous solution is only a special case, a closed form solution generally does not exist. Therefore iterative optimization algorithms such as coordinate descent are used. The algorithm works by these steps:

1. Initialize an estimate for β with Ordinary Least Squares coefficients
2. Cycle through each coordinate j, updating β_j by minimizing $g(\beta)$ with respect to β_j
3. Apply the soft-thresholding operator to the partial residuals
4. Iterate until convergence

λ is chosen the same way as ridge regression through an iterative validating selection algorithm[14].

4 Linear Discriminant Analysis

4.1 The Setup

Start with a classification problem that has K classes. We have n observations(data points), where each observation is a d-dimensional data vector $\vec{x_i}$ and an associated class label $y_i \in \{1, 2, \ldots, K\}$. For each class k, define a class mean $\vec{u}_k$ and assume that the data within each class are drawn from a multivariate normal distribution with common covariance matrix Σ[15].

4.2 The Model

4.2.1 Calculate Scatter Matrices

We calculate the within-class scatter matrix (S_W) and the between-class scatter matrix (S_B). S_W measures the spread of samples within each class—that is, how much the data points in a class deviate from their class mean $\boldsymbol{\mu}_k$. S_B measures the scatter of class means relative to the overall mean $\boldsymbol{\mu}$, showing how separated the class means are.

14. Tilevak 2022.
15. IBM 2023.

$$S_W = \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} (\mathbf{x} - \boldsymbol{\mu}_k)(\mathbf{x} - \boldsymbol{\mu}_k)^T$$

$$S_B = \sum_{k=1}^{K} n_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T$$

Here:

- C_k represents the set of observations in class k.
- n_k is the number of observations in class k.
- $\boldsymbol{\mu}$ is the overall mean of all observations[16].

4.2.2 Formulate the Optimization Problem

LDA's goal is to find a projection vector w that maximizes the ratio of the between-class scatter to the within-class scatter. This is known as the Fisher criterion.

$$J(w) = \frac{w^T S_B w}{w^T S_w w}$$

Maximizing this ratio means finding a direction where class means are far apart(large numerator) and data points within classes are tightly clustered(small denominator)

4.3 Constraint

Directly maximizing the Fisher criterion is very challenging so we impose the constraint:

$$w^T S_w w = 1$$

Under this constraint, maximizing $J(w)$ is equivalent to maximizing $w^T S_B w$. With this constraint, we can use a Lagrange multiplier

16. 2023.

4.4 Lagrangian

The Lagrangian function that uses this constraint is:

$$L(w, \lambda) = w^T S_B w - \lambda(w^T S_W w - 1)$$

where λ is a Lagrange multiplier[17].

4.5 Derivation

Remember that:

$$f(w) = w^T A w$$

$$\nabla_w f(w) = 2Aw$$

when A is a symmetrical matrix. S_W and S_B are both symmetrical by definition. Therefore,

$$\frac{\partial L}{\partial w} = 2S_B w - 2\lambda S_W w = 0$$

Simplifying leads to

$$S_B w = \lambda S_W w$$

4.6 Eigenvalue problem

Assuming S_W is invertible, we rewrite the previous step as:

$$S_W^{-1} S_B w = \lambda w$$

This should look very familiar if you are experienced with eigenvalues. Rewrite $S_W^{-1} S_B$ as U.

$$Uw = \lambda w$$

Which is exactly the definition of eigenvectors. Simply find the greatest eigenvalue and its corresponding eigenvector. Project points onto this vector and project means on to this vector[18].

17. Gábor Balázs. 2024. *How can I use Lagrangian Multipliers to maximize a General Rayleigh Quotient for Linear Discriminant Analysis.* Forum, February. Accessed April 1, 2025. https://math.stackexchange.com/questions/4843451/how-can-i-use-lagrangian-multipliers-to-maximize-a-general-rayleigh-quotient-for.

18. 2023.

4.7 Use

For each data point, project it onto the found Eigenvector and find the projected mean that it is closest to. The corresponding class mean is the class that the model predicts the point belongs to.

5 Gaussian Naive Bayes' Classifier

5.1 The Setup

Naive Bayes' classifiers have three components. The first component is, of course, Bayes' Rule. The second component is the "naive" assumption about the data. The final component is the PDF/PMF that is used for calculating probabilities.

Bayes' Rule

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Sometimes also written, in the more readable way,

$$\text{The Posterior Probability} = \frac{\text{Prior Probability} \cdot \text{Likelihood}}{\text{Evidence}}$$

The naive assumption is that all features of the data are independent conditioned on the class. So if the data we were using was flowers, the class could be rose and the features could be pedal size, stem size, leaf size, etc.. It is reasonable to assume that the size of a flowers petal's are unaffected by the size of the stem given it is a rose or poppy or some other type of flower. This is the Naive assumption[19]

$$P(x_i|x_1, \ldots, x_n, C_k) = P(x_i|C_k)$$

Because we are looking at a Gaussian Naive Bayes' classifier instead of a Multinomial, Bernoulli, Semi-supervised parameter estimation, or other Naive Bayes' classifier, we use the Gaussian Distribution. Therefore the PDF is

$$P(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

19. Weinberger 2018.

Where σ is the standard deviation, σ^2 is the variance, μ is the mean, and x is the input[20].

5.2 The Model

First define k classes, $C_1, C_2, \ldots C_k$. Each data point, of n data points, has features $x_1, x_2, \ldots, x_m$ and a class C. All features x are continuous variables and assumed to be distributed on a Gaussian Distribution dependent on their class. So, the distribution of pedal sizes of a rose is unrelated to that of a poppy. So for each feature of each class, define a normal distribution

$$J_{km} \sim \mathcal{N}(\mu_{km},\, \sigma_{km}^2)\,.$$

Then go through all training data to estimate these distributions. For each class and feature, calculate the mean

$$\mu_{km} = \frac{1}{p_k}\sum_{i=1}^{p_k} d_{kim}$$

where p_k is the number of data points in class C_k in the training data and d_{ki} is the ith data point in class k. Then the variance is

$$\sigma_{km}^2 = \frac{1}{p_k}\sum_{i=1}^{p_k}(\mu_{km} - d_{kmi})^2$$

Once these values are recorded, we need to calculate the probability of a data point being in a class. The data point can be represented as all of it's features.

$$P(C_k|x_1, x_2, \ldots, x_m) = \frac{P(C_k)P(x_1, x_2, \ldots, x_m|C_k)}{P(x_1, x_2, \ldots, x_m)}$$

Since the denominator is not influenced by C_k and stays constant, it does not help us determine what class a data point is in so we can ignore it and only look at the numerator[21].

$$P(C_k)P(x_1, x_2, \ldots, x_m|C_k)$$

20. 2018.
21. *Id.*

Because of the joint probability model

$$P(C_k)P(x_1, x_2, \ldots, x_m|C_k) = P(x_1, x_2, \ldots, x_m, C_k)$$

Then

$$P(x_1, x_2, \ldots, x_m, C_k) = P(x_1|x_2, \ldots, x_m, C_k)P(x_2, \ldots, x_m, C_k)$$

Continuing this

$$P(x_1|x_2, \ldots, x_m, C_k)P(x_2, \ldots, x_m, C_k)$$
$$= P(x_1|x_2, \ldots, x_m, C_k)P(x_2|x_3, \ldots, x_m, C_k)P(x_3, \ldots, x_m, C_k)$$
$$= P(x_1|x_2, \ldots, x_m, C_k)P(x_2|x_3, \ldots, x_m, C_k) \ldots P(x_{m-1}|x_m, C_k)P(x_m|C_k)P(C_k)$$

Now we use the naive assumption from The Setup.

$$= P(x_1|C_k)P(x_2|C_k) \ldots P(x_{m-1}|C_k)P(x_m|C_k)P(C_k)$$

These probabilities are all easy to calculate given the assumption of Normally distributed features. The first probability to tackle is $P(C_k)$. This can be estimated simply by

$$\frac{\text{Data points that are of class K}}{\text{Total data points}}$$

To calculate the probabilities of $P(x_i|C_k)$ we use the estimated distribution from before. So

$$P(x_m|C_k) = \text{normalPDF}(x_m, \mu_{km}, \sigma^2_{km})^{22}$$

We then calculate this for every class k for this data point and the class with highest resulting probability is the presumed class

$$P(x_1|C_k)P(x_2|C_k) \ldots P(x_{m-1}|C_k)P(x_m|C_k)P(C_k)$$

6 Logistic Regression

6.1 The Setup

Logistic regression is a classification method that models the probability of a binary outcome using the logistic function. Instead of assuming normally distributed

features, as with the Gaussian Naive Bayes' Classifier, the model applies a transform to a linear transformation of the features. The core of the model, of course, is the logistic/sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Where z is defined as a linear combination of m input features

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

And, β_0 is the intercept, sometimes also called the bias term. $\beta_1, \beta_2, \ldots, \beta_m$ correspond to features $x_1, x_2, \ldots, x_m$ respectively[23]. The function transforms any real-valued number into a value between 0 and 1, interpreted as the probability of the data point belonging to a particular class. Since logistic regression is purely focused on modeling the probability via the logistic function, there is no requirement to assume the independence among features.

6.2 The Model

In a binary classification problem with two classes, C_0 and C_1, let y be a response variable where $y = 1$ corresponds to C_1 and $y = 0$ corresponds to C_0. The logistic regression model expresses the probability that a data point belongs to C_1 with features $x_1, x_2, \ldots, x_m$ as

$$P(y = 1|x_1, x_2, \ldots, x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}}$$

Since the two classes are complementary, a data point has to be one or the other, the probability that that data point belongs to C_0 is

$$P(y = 0|x_1, x_2, \ldots, x_m) = 1 - P(y = 1|x_1, x_2, \ldots, x_m)$$

23. Daniel Jurafsky and James H. Martin. 2025. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition with language models.* 3rd. Online manuscript released January 12, 2025. Stanford. https://web.stanford.edu/~jurafsky/slp3/.

The parameters $\beta_1, \beta_2, \ldots, \beta_m$ are estimated from the training data using the likelihood function. For a training set with n samples $\{x^{(i)}, y^{(i)}\}_{i=1}^{n}$ the likelihood function is

$$L(\beta) = \prod_{i=1}^{n} [P(y^{(i)}|x^{(i)})]^{y^{(i)}} [1 - P(y^{(i)}|x^{(i)})]^{1-y^{(i)}}$$

This function is a way to measure how effective the β terms are. There are two main terms in this function. If the true class is C_1 then the first term is multiplied in the product. If the true class is C_1 then the second term is multiplied in the product where each term just represents the confidence the model had for that class. Taking the natural log makes the function a summation instead of a product, this is called the log-likelihood function[24].

$$\ell(\beta) = \sum_{i=1}^{n} \left[y^{(i)} \log P(y^{(i)}|x^{(i)}) + \left(1 - y^{(i)}\right) \log \left(1 - P(y^{(i)}|x^{(i)})\right) \right]$$

Then the log-likelihood function is maximized using gradient descent or the Newton-Raphson algorithm. New data points / testing data points are classified by computing $P(y = 1|x_1, x_2, \ldots, x_m)$.

$$\text{Class} = \begin{cases} C_1 & \text{if } P(y = 1|x_1, x_2, \ldots, x_m) > 0.5, \\ C_0 & \text{otherwise.} \end{cases}$$

24. Jurafsky and Martin 2025.

Part II
Experiment

7 Overview of the data

To test the models described in this paper, two extremely popular data sets were tested. We also will alter the data in ways that will change the effectiveness of the models. The alteration of data will **not** remove any information from training or testing. The alteration will be used to test the effectiveness of the models if a researcher does not follow appropriate steps in collecting or cleaning data.

7.1 Breast Cancer Data

The breast cancer data set is the data set used to test the classification models. Having only two classes, malignant and benign, it makes implementing the models much easier. It also shows a very practical use of machine learning to help cancer screening. The data set consists of 569 entries, which is less than ideal for training a model. There are 212 malignant data points and 357 benign data points. Each data point has 30 features[25].

7.2 California Housing Data

The California housing data set is the data set used to test the regression models. The data is ideal to test with linear regression because it is fair to assume at least some features will scale linearly with house price. For example square feet of a home · price per square foot is commonly used to value homes. The data set consists of $20,640$ entries and 9 features per data point. Instead of collecting data on individual homes, the data points represent blocks of adjacent homes. So instead of bedrooms in a home, it is average bedrooms per home in a block. This helps eliminate non-quantitative features like house nice a home is painted[26].

25. scikit-learn developers. 2025b. *Sklearn.datasets.load_breast_cancer.* https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html. Accessed April 6, 2025.

26. scikit-learn developers. 2025a. *Sklearn.datasets.fetch_california_housing.* https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html.

8 Heuristic Model Limitations

The fundamental models of machine learning, the ones focused on in this paper, have lower requirements of data size than a massive neural network. They can achieve stable and interpretive results at order of magnitudes lower than more complicated models that make less assumptions about the data.

8.1 Logistic Regression

Although logistic regression is an incredibly powerful and popular model, requires 10 events per variable (EPV)[27]. So, in a dataset that has 30 features, 300 samples should be enough.

8.2 Naive Bayes' Classifier

Unlike logistic regression, Naive Bayes' Classifiers have no generally accepted EPV. However, the strong assumptions in the model, even when not true, generally result in a classifier that works well[28].

8.3 Discriminant Analysis

Discriminant analysis does not have an a general EPV. Instead it has an EPV of 3 for each group being separated[29]. So there would need to be 90 samples in each group for a dataset with 30 features. So, the model is reasonable to apply to this dataset.

8.4 Ordinary Least Squares

Ordinary Least Squares, and other regression models, have a general rule of thumb for 10 samples per feature. However, some models have found that only 2 samples

Accessed April 6, 2025.

27. P. Peduzzi et al. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 49 (12): 1373–1379. https://doi.org/10.1016/s0895-4356(96)00236-3.

28. Kevin P. Murphy. 2012. *Machine learning: a probabilistic perspective.* Cambridge, MA: MIT Press. ISBN: 9780262018029.

29. Peter C. Austin and Ewout W. Steyerberg. 2017. Events per variable (epv) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Epub 2014 Nov 19, *Statistical Methods in Medical Research* 26 (2): 796–808. https://doi.org/10.1177/0962280214558972.

	Logistic Regression	Discriminant Analysis	Bayes' Classifier
Accuracy %	98.25	95.32	93.57
Sensitivity %	98.41	90.48	90.48
Specificity %	98.15	98.15	95.37
Validity %	96.56	88.62	85.85
Runtime (seconds)	0.023	0.015	0.015

Table 1: Unaltered Data Testing Results (Classification)

are needed per feature for an accurate regression[30]. There is a wide range of "acceptable" EPVs but, it is important to look at model performance instead of having a set EPV for Ordinary Least Squares.

8.5 Ridge and LASSO Regression

Ridge and LASSO regression generally require less samples than Ordinary least squares since they are more robust to overfitting. However, a general rule of thumb could be the same at 10 samples per feature. Again though, checking model performance is extremely important in determining if the model is trained on enough data.

9 Comparison of unaltered data

Source Code and Figures

9.1 Classification

Refer to Table 3 for figures. In the measure of accuracy, logistic regression was best, followed by LDA, followed by GNBC. All models had above 90% accuracy which is very good. However, logistic regression outperformed the others significantly. Although logistic regression performed the best, it is also the most energy intensive. Instead of the closed form solutions of GNBC and LDA, logistic regression has to have it's coefficients, weights and bias, estimated. This estimation requires more computing power. Discriminant analysis was the next best model and it has a

30. Peter C. Austin and Ewout W. Steyerberg. 2015. The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology* 68 (6): 627–636. ISSN: 0895-4356. https://doi.org/10.1016/j.jclinepi.2014.12.014. https://www.jclinepi.com/article/S0895-4356(15)00014-1/fulltext.

closed form solution. Discriminant analysis had 0.015 seconds of runtime and GNBC also had 0.015 seconds of runtime. Logistic regression had 0.023 seconds of runtime, that is more than a 50% increase in runtime. While all three models runtime is not very long, if they were used in an application where the model was retrained very often, it could become expensive to use logistic regression compared to the other models.

It is important to look at more than just accuracy, though, since that does not paint a complete picture of a model's performance. Sensitivity is an incredibly important benchmark for cancer screening. This is because sensitivity relates the true reported positives to the total amount of positives. This is important in cancer screening because it is more important to minimize false negatives than false positives. If a doctor screens a patient for cancer using a model and it reports negative when the patient has cancer, it is much more detrimental than reporting that the patient has cancer when they don't. This is because further tests will be conducted to make sure a person has cancer if the model reports it but, no further tests will be conducted if the model reports a person does not. Again, logistic regression is the best performing model. LDA and GNBC had identical and less than ideal results.

Specificity is the inverse of sensitivity. Specificity relates the true reported negatives to the total amount of negatives. This is less important in cancer screening but, it is not ideal to send a patient to get more invasive tests for cancer unless it is necessary. It is both expensive financially and personally to a patient. Logistic regression performed by far the best in this benchmark while GNBC and LDA performed identically.

Validity is just a balance of specificity and sensitivity. It is important to examine when sensitivity and specificity are approximately of the same importance. Although it is not as important as sensitivity in the case of cancer screening, it does seem to order our models in a way that makes sense. Logistic regression being the most effective then discriminant analysis being worse and then Gaussian Naive Bayes' Classifier being marginally worse than discriminant analysis.

9.2 Prediction

	Ordinary Least Squares	Ridge Regression	LASSO Regression
R^2	0.5833	0.6001	0.6006
Runtime (seconds)	0.005	0.066	5.606

Table 2: Unaltered Validation Results(Prediction)

Refer to Table 2 for figures. For prediction models, there are 2 benchmarks to observe, one is Mean Squared Error and the other is R^2. However, R^2 is actually just a normalized Mean Squared Error which makes it easier to compare across data sets. It represents how much of the response variable is able to be predicted with the input variables. So a value of 0.5 would mean that 50% of a response variable is predicted by the input variables. A value of 0 would mean that the response variable is not at all correlated with the input variables.

All the models performed similarly, although the more complicated regressions performed better than ordinary least squares. As talked about in section 3.1, the weakness of ordinary least squares is that it can over fit to training data. This leads to the model "learning" the noise and the patterns instead of just the patterns. Leading it to perform worse on testing data that has different noise.

Ridge regression and LASSO regression had very similar R^2 values on the validation set although LASSO regression was the best. However, R^2 does not paint a full picture, LASSO regression has to estimate β for many λ values while ridge regression has a closed form solution for β for each λ. Ordinary least squares, on the other hand, has just one closed form solution, making it much faster to run. Lasso took over 1000 times longer to run than ordinary least squares. Ridge regression was more than 10 times slower to run than ordinary least squares. The nuance of runtime makes it a lot less obvious which model to choose.

10 Comparison of altered data

10.1 Classification

	Logistic Regression	Discriminant Analysis	Bayes' Classifier
Accuracy %	97.66	96.49	94.15
Sensitivity %	98.41	95.24	88.89
Specificity %	97.22	97.22	97.22
Validity %	95.63	92.46	86.11

Table 3: Altered Data Testing Results (Classification)

To alter the data for classification, we altered only the training data. We used a technique called SMOTE (Synthetic Minority Over-sampling Technique) to generate synthetic data for the malignant class. This technique works by getting data points in the minority class and then averaging their features with the features of their closest neighbors, creating realistic data. This situation could occur when

doctors report all malignant cases but, only some benign cases. This changes the distribution of malignant and benign in the training data. However, a doctor is likely to use the model on all tumors to determine if the tumor is malignant. This creates a discrepancy between training and testing data distributions.

Since the training data has changed a bit, it is important to notice large changes in benchmarks and ignore smaller ones. GNBC had one more false positive and 2 less false negatives with the altered data compared to the unaltered data. The model has become a lot more likely to predict a positive, or malignant, result. Since GNBC is a generative model, unlike the other two discriminative models, it attempts to learn something about the classes themselves, including the likelihood of a class in the distribution[31]. This means it performance can be drastically altered when the training data and testing data have different distributions of classes.

Discriminant analysis had much better performance on the altered data since it was likely able to calculate the class means, along with having more separation between classes. Logistic regression had only 1 more false negative than the unaltered data. Discriminative models try to figure out differences between classes instead of trying to figure out what a class is. For example, in a data set of cats and dogs, if all the dogs wear collars and none of the cats do, that is sufficient for discriminative models. Whereas a generative model will try to figure out what a cat looks like and what a dog looks like instead of finding discriminative features.

10.2 Prediction

Note: Ordinary least squares code had to be modified because SKLearn uses singular value decomposition under the hood and is able to find a solution even with a singular matrix[32]. Therefore, we took a more bare-bones approach with numpy and direct matrix calculations. Tested λ ranges for lasso and ridge regression were condensed closer to 1. This is because the same λs should be tested for ridge and lasso. Lasso would not converge for smaller λ.

	Ordinary Least Squares	Ridge Regression	Lasso Regression
R^2	-357.85	0.58	0.60

Table 4: Altered Data(Prediction)

31. 2025.

32. SciPy Developers. 2023. *Scipy/linalg/_basic.py at commit ef7a30c56dbaddea1688fd2dfcb56022a5d3bb07.* https : / / github . com / scipy / scipy / blob / ef7a30c56dbaddea1688fd2dfcb56022a5d3bb07 / scipy / linalg / _basic . py # L1326. Accessed: 2025-04-06.

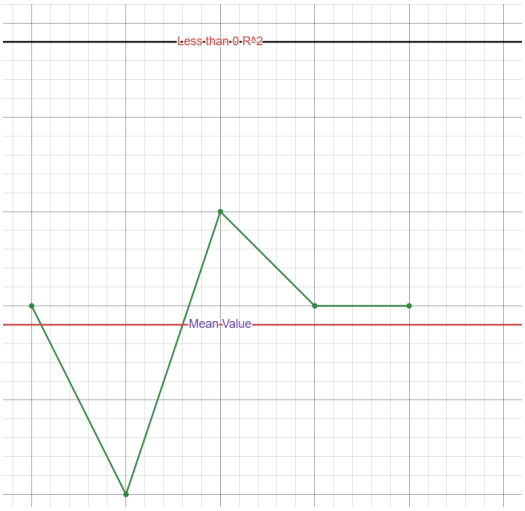

Figure 4: Less than 0 R^2 Value

To alter the data for prediction, we added a column to X which was $\vec{x_3} \cdot 2$. This made a linearly dependent column. Technically, in programming, floating point operations have error so the new column is not fully linearly dependent. It is very close and shows the power of ridge and lasso regression. One way in which this data could have been created by a rookie data researcher is when he or she would like to "add" more information to a data set after collecting the data. He or she decides that the third column, "average rooms", is more important and doubles the column and adds it to the end of the data. This shows a fundamental lack of understanding of how linear regression works. The results of his or her analysis now seems to be a mathematical or logical bug in the program.

Although it is normally impossible for an R^2 value to be negative, it is possible if the model is extremely poorly fit. A 0 R^2 value is possible by estimating the data with the mean value. This means that the ordinary least squares model is performing worse than just taking the average of the data, see Figure 4. Ridge and Lasso regression also perform worse than the unaltered data although it is not very significant and could be due also in part to the change in the tested λ range.

This shows the danger of having a model that becomes so overfit, that it is underfit to the testing data.

11 Conclusion

In summary, this thesis provides an in-depth analysis of fundamental machine learning models for both prediction and classification. It highlights their respective strengths and vulnerabilities. The comparative study of logistic regression, discriminant analysis, and Bayes' classifier on the breast cancer dataset illustrated that while logistic regression achieved higher benchmark, it also demanded higher computational resources. The higher resource usage coming from its iterative coefficient estimation. This trade-off demonstrates the importance in selection of models that balance performance and computational efficiency a necessary task.

For prediction models, the evaluation of ordinary least squares, ridge, and lasso regression on the California housing data illustrated the risk in overfitting. The experiments demonstrated that while ordinary least squares can perform adequately on well-behaved data, small issues with the data (near collinearity) can severely degrade its performance. In contrast, ridge and lasso regression, through their regularization techniques, proved more robust in data manipulation. However, this robustness came at the cost of increased computational time, especially for lasso regression.

The intentional alteration of datasets to simulate real-world data collection/analysis challenges demonstrated the vital need for careful data preprocessing and collection. For classification models, this came from using the same distribution of data for training and testing. For prediction, this came at not collecting collinear columns when using ordinary least squares regression. These observed challenges in model performance serve as a cautionary example for researchers and data collectors alike.

Overall, this work reinforces that the choice of model and preprocessing methods must be guided by the characteristics of the dataset and requirements of the application. As machine learning continues evolving, so too must the methodology for model selection, validation, and interpretation to ensure the derived insights are accurate and actionable.

References

Amazon. *What is overfitting?* Accessed April 1, 2025. https://aws.amazon.com/what-is/overfitting/.

Austin, Peter C., and Ewout W. Steyerberg. 2015. The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology* 68 (6): 627–636. ISSN: 0895-4356. https://doi.org/10.1016/j.jclinepi.2014.12.014. https://www.jclinepi.com/article/S0895-4356(15)00014-1/fulltext.

———. 2017. Events per variable (epv) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. Epub 2014 Nov 19, *Statistical Methods in Medical Research* 26 (2): 796–808. https://doi.org/10.1177/0962280214558972.

Balázs, Gábor. 2024. *How can I use Lagrangian Multipliers to maximize a General Rayleigh Quotient for Linear Discriminant Analysis.* Forum, February. Accessed April 1, 2025. https://math.stackexchange.com/questions/4843451/how-can-i-use-lagrangian-multipliers-to-maximize-a-general-rayleigh-quotient-for.

Bellhouse, D. R. *The reverend thomas bayes, frs: a biography to celebrate the tercentenary of his birth.* Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University. Accessed March 6, 2025. This source provides information on Thomas Bayes and his contributions to mathematics and statistics, emphasizing the significance of Bayes' rule. https://biostat.jhsph.edu/courses/bio621/misc/bayesbiog.pdf.

Brown, Sara. 2021. *Machine learning, explained.* MIT Management Sloan School. Last modified April 2021. Accessed December 10, 2024. This source explains machine learning concepts, applications, and types. Published by MIT, it provides a strong foundation for understanding the field. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

Buttari, Alfredo, Victor Eijkhout, Julien Langou, and Salvatore Filippone. 2007. Performance optimization and modeling of blocked sparse kernels. This peer-reviewed source discusses optimizing block sparse kernels with performance evaluation. It contains graphs and implementation results, useful for classification algorithm optimization. *International Journal of High Performance Computing Applications* 21 (4): 467+.

Carnahan, Brian, Gerard Meyer, and Lois-Ann Kuntz. 2003. Comparing statistical and machine learning classifiers: alternatives for predictive modeling in human factors research. This peer-reviewed source compares traditional statistical classifiers with machine learning approaches, such as decision trees and genetic programming, in human performance applications. *Human Factors* 45 (3): 408+.

Chen, Xinye, and Stefan Güttel. 2024. Fast and exact fixed-radius neighbor search based on sorting. This source introduces a sorting-based method for K-nearest neighbors (SNN) to reduce query times. The authors have strong backgrounds in computational mathematics. *PeerJ Computer Science* 10:e1929.

Cristianini, Nello, and Bernhard Scholkopf. 2002. Support vector machines and kernel methods: the new generation of learning machines. This peer-reviewed article discusses the transition from support vector machines to kernel methods, providing insights into their development and applications. *AI Magazine* 23 (3): 31+.

Deisenroth, Marc Peter, A. Aldo Faisal, and Cheng Soon Ong. 2020. *Mathematics for machine learning.* This widely recognized book covers mathematical foundations of machine learning, structured by relevant mathematical fields leading to real-world applications. Cambridge University Press.

Djunaidy, Arif, and Nisrina Fadhilah Fano. 2024. Development of customer review ranking model considering product and service aspects using random forest regression method. This peer-reviewed source presents regression-based customer review ranking models, with comparisons to existing algorithms. *KSII Transactions on Internet and Information Systems* 18 (8): 2137+.

Domingos, Pedro. 2012. A few useful things to know about machine learning. This highly cited article provides fundamental insights into machine learning concepts, terminology, model evaluation, and common pitfalls. *Communications of the ACM* 55 (10): 78–87. https://doi.org/10.1145/2347736.2347755.

IBM. 2023. *What is linear discriminant analysis (lda)?* https://www.ibm.com/think/topics/linear-discriminant-analysis. Last modified November 27, 2023. Accessed March 6, 2025. This source provides an informative introduction to LDA and its mathematical foundations.

———. 2024. *What is lasso regression?,* January. Accessed April 1, 2025. https://www.ibm.com/think/topics/lasso-regression.

Jurafsky, Daniel, and James H. Martin. 2025. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition with language models.* 3rd. Online manuscript released January 12, 2025. Stanford. https://web.stanford.edu/~jurafsky/slp3/.

Mu, Ruihui, and Xiaoqin Zeng. 2019. A review of deep learning research. This paper reviews deep learning applications, models, and optimization techniques, authored by well-published researchers. *KSII Transactions on Internet and Information Systems* 13 (4): 1738+.

Murel, Jacob, and Eda Kavlakoglu. 2023. *What is ridge regression?* Accessed November 21, 2023. https://www.ibm.com/think/topics/ridge-regression.

Murphy, Kevin P. 2012. *Machine learning: a probabilistic perspective.* Cambridge, MA: MIT Press. ISBN: 9780262018029.

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. 1996. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 49 (12): 1373–1379. https://doi.org/10.1016/s0895-4356(96)00236-3.

Rong, Jian-Ying, and Xu-Qing Liu. 2024. Hybrid principal component regression estimation in linear regression. This source explores hybrid regression models and their performance, with visual comparisons demonstrating improvements over traditional approaches. *Electronic Research Archive* 32 (6): 3758+.

scikit-learn developers. 2025a. *Sklearn.datasets.fetch_california_housing.* https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html. Accessed April 6, 2025.

———. 2025b. *Sklearn.datasets.load_breast_cancer.* https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html. Accessed April 6, 2025.

SciPy Developers. 2023. *Scipy/linalg/_basic.py at commit ef7a30c56dbaddea1688fd2dfcb56022a5d3bb07.* https://github.com/scipy/scipy/blob/ef7a30c56dbaddea1688fd2dfcb56022a5d3bb07/scipy/linalg/_basic.py#L1326. Accessed: 2025-04-06.

Shalizi, Cosma. 2015. *Simple linear regression in matrix format.* Carnegie Mellon University Statistics and Data Science, Carnegie Mellon University. Last modified October 13, 2015. Accessed March 6, 2025. This lecture summary explains least squares regression using matrix notation. https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/13/lecture-13.pdf.

Tabachnick, Barbara G, and Linda S Fidell. 2013. *Using multivariate statistics: pearson new international edition* [in en]. 6th ed. Pearson custom library. London, England: Pearson Education, July.

Tifenbach, Ryan M. 2011. On an svd-based algorithm for identifying meta-stable states of markov chains. This source discusses singular value decomposition (SVD) for identifying meta-stable states in Markov chains. *Electronic Transactions on Numerical Analysis* 38:17+.

Tilevak, Andreas. 2022. *Lasso regression - explained,* July. Accessed April 1, 2025. https://www.youtube.com/watch?v=bPFjfZWWQO0.

Weinberger, Kilian. 2018. *Bayes classifier and naive bayes,* July. Accessed April 1, 2025. https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote05.html.

Yong, York. *Introduction to naive bayes algorithm - gaussian and multinomial variants.* Accessed April 1, 2025. https://www.kaggle.com/discussions/general/468420.

Zhang, Jianjun, and Benedetta Morini. 2013. Solving regularized linear least-squares problems by the alternating direction method with applications to image restoration. This source discusses least-squares problems and their applications, useful for regression modeling. *Electronic Transactions on Numerical Analysis* 40:356+.