introduction to machine learning

DR.Amiri

electrical engineering department

Ahmadreza Majlesara 400101861

assignment 2

July 11, 2024

introduction to machine learning

assignment 2

Ahmadreza Majlesara 400101861

Easy peasy lemon squeezy

Suppose that we have collected the data of a group of students of ML course so that this data includes two features. The first characteristic, x_1, is equal to the total number of hours that the student has practiced ML. The second characteristic, x_2, is the student's total grade point average (GPA) before taking the ML course. Now, we are going to use a logistic regression model to predict the probability that a student will get a score of 20 in this course. After learning based on these data, the obtained coefficients are equal to $\beta_0 = -5$, $\beta_1 = 0.1$ and $\beta_2 = 0.25$.

part 1

Calculate the probability that a student with 80 hours of study and an GPA of 18 can get a score of 20 in this course.

solution

we can model this problem as a binary logistic regression problem. we know that for a binary logistic regression problem we have:

$$P(y=1|x;\theta) = \sigma(\alpha)$$

where σ is the sigmoid function and $\alpha = w^T x + b$ where w is the weight vector and b is the bias and x is the input vector. now if we say y is the probability of getting a score of 20 in this course we can write:

$$P(y=1|x;\theta) = \sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

$$x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 80 \\ 18 \end{bmatrix}$$

$$w = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} -5 \\ 0.1 \\ 0.25 \end{bmatrix}$$

$$P(y=1|x;\theta) = \sigma(-5 + 0.1 \times 80 + 0.25 \times 18) = \sigma(7.5) = \frac{1}{1+e^{-7.5}} = 0.9994$$

part 2

Consider another student who has a GPA of 16 and wants to get a grade of 20 in this course. According to the model that we have taught, how many hours should he practice in order to achieve this grade with a 90% probability?

solution

we use the previous model in this part too. we have:

$$x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ x_1 \\ 16 \end{bmatrix}$$

$$w = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} -5 \\ 0.1 \\ 0.25 \end{bmatrix}$$

$$P(y=1|x;\theta) = \sigma(\alpha) \geq 0.9 \Rightarrow \frac{1}{1+e^{-\alpha}} \geq 0.9$$

$$\Rightarrow e^{-\alpha} \leq \frac{1}{9} \Rightarrow \alpha \geq ln(9)$$

$$\Rightarrow \beta_0 + \beta_1 x_1 + \beta_2 x_2 \geq ln(9)$$

$$\Rightarrow -5 + 0.1x_1 + 0.25 \times 16 \geq ln(9)$$

$$\Rightarrow 0.1x_1 \geq ln(9) + 1 \Rightarrow x_1 \geq \frac{ln(9)+1}{0.1} = 31.97$$

Multi-class Logistic Regression

One way to extend logistic regression to multi-class (say K class labels) setting is to consider $(K-1)$ sets of weight vectors and define

$$P(Y=y_k \mid X)=\frac{\exp\left(w_{k0}+\sum_{i=1}^{d} w_{ki}x_i\right)}{\sum_{j=1}^{K-1}\exp\left(w_{j0}+\sum_{i=1}^{d} w_{ji}x_i\right)}, \quad \text{for } k=1,\ldots,K-1$$

part 1

What does this model imply for $P(Y=y_k \mid X)$?

solution

$$P(Y=y_k \mid X)=1-\sum_{k=1}^{K-1} P(Y=y_k \mid X)=1-\sum_{k=1}^{K-1}\alpha\exp\left(w_{k0}+\sum_{i=1}^{d} w_{ki}x_i\right)$$

$$\Rightarrow \alpha=\frac{1}{1+\sum_{k=1}^{K-1}\exp\left(w_{k0}+\sum_{i=1}^{d} w_{ki}x_i\right)}$$

so the model implies that the probability of the Kth class is the complement of the sum of the probabilities of the other classes means that:

$$P(Y=y_K \mid X)=\alpha\exp\left(w_{K0}+\sum_{i=1}^{d} w_{Ki}x_i\right), \quad \text{for } k=1,2,\ldots,K-1$$

$$P(Y=y_K \mid X)=\alpha, \quad \text{for } k=K$$

part 2

What would be the classification rule in this case?

solution

the classification rule in this case is to choose the class with the highest probability. for example for class j we choose the class with the highest $P(Y=y_j \mid x)$

part 3

Draw a set of training data with three labels and the decision boundary resulting from a multi-class logistic regression The boundary does not need to be quantitatively correct but should qualitatively depict how a typical boundary from multi-class logistic regression would look like.

solution

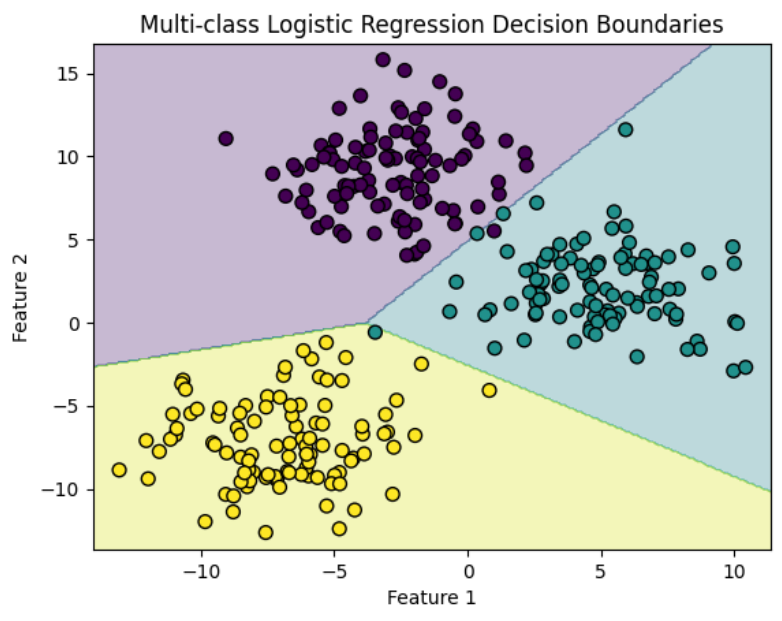

Figure 1. Multi-class logistic regression decision boundary

part 4

Find log-likelihood if N data sample: $X = \{(x_i, y_i)\}_{i=1}^{N}$ is observed

solution

as we proved in part 1 the probability of the Kth class is:

$$P(Y = y_K \mid X) = \alpha \exp\left(w_{K0} + \sum_{i=1}^{d} w_{Ki} x_i\right), \quad \text{for } k = 1, 2, \ldots, K-1$$

$$P(Y = y_K \mid X) = \alpha, \quad \text{for } k = K$$

wich α is:

$$\alpha = \frac{1}{1 + \sum_{k=1}^{K-1} \exp\left(w_{k0} + \sum_{i=1}^{d} w_{ki} x_i\right)}$$

so the log-likelihood of the data is:

$$\mathscr{L}(w) = \sum_{i=1}^{N} \log P(y_i \mid x_i; w)$$

$$= \sum_{i=1}^{N} \log \left(\frac{\exp\left(w_{y_i 0} + \sum_{j=1}^{d} w_{y_i j} x_{ij}\right)}{\sum_{k=1}^{K-1} \exp\left(w_{k0} + \sum_{j=1}^{d} w_{kj} x_{ij}\right)} \right)$$

part 5

Now suppose that we add a L2 regularizing term to this objective function:

$$f(X) = \log(P(Y \mid X)) - \lambda \sum_{k=1}^{K} \|w_k\|^2$$

Now Calculate the gradient of the function $f(X)$ with respect to w.

solution

$$f(W) = \mathscr{L}(w) - \lambda \sum_{k=1}^{K} \|w_k\|^2$$

$$\Rightarrow \nabla f(W) = \nabla \mathscr{L}(w) - \lambda \nabla \sum_{k=1}^{K} \|w_k\|^2$$

$$= \nabla \mathscr{L}(w) - 2\lambda w$$

$$\Rightarrow \nabla_w f(W) = \sum_{i=1}^{N} \left[\frac{x_i \exp(w_k^T x_i)}{1 + \sum_{j=1}^{K-1} \exp(w_j^T x_i)} \right] - 2w_k$$

part 6

Write update rule for the dataset X according to the answer of the previous part and simplify as much as possible and explain what happens to w at each iteration

solution

by gradient descent we have:

$$w_k^{(t+1)} = w_k^{(t)} - \eta \nabla_w f(W)$$

$$= w_k^{(t)} - \eta \left[\sum_{i=1}^{N} \left[\frac{x_i \exp(w_k^{(t)^T} x_i)}{1 + \sum_{j=1}^{K-1} \exp(w_j^{(t)^T} x_i)} \right] - 2w_k^{(t)} \right]$$

Overfitting and Regularized Logistic Regression

Part 1

Plot the sigmoid function $\frac{1}{1+e^{-uX}}$ vs. $X \in \mathbb{R}$ for increasing weight $u \in \{1, 5, 100\}$. A qualitative sketch is enough. Use these plots to argue why a solution with large weights can cause logistic regression to overfit.

solution

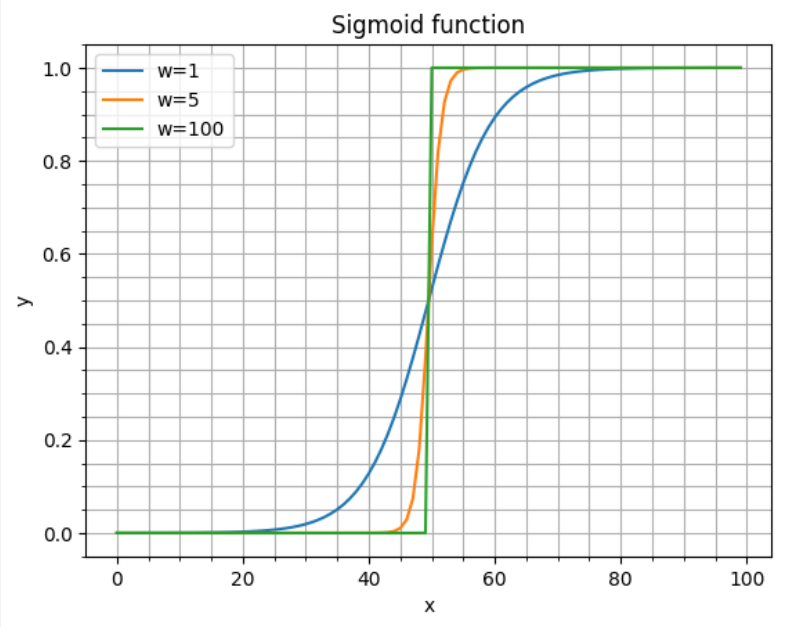

Figure 2. Sigmoid function for increasing weight w

As the weight w increases, the sigmoid function becomes steeper, which means that the output of the logistic regression model will be more sensitive to small changes in the input. This can cause the model to fit the training data too closely, capturing noise in the data rather than the underlying pattern. This is known as overfitting, and it can lead to poor generalization performance on new, unseen data.

Part 2

To prevent overfitting, we want the weights to be small. To achieve this, instead of maximum conditional likelihood estimation (M(C)LE) for logistic regression:

$$\max_{w_0,\ldots,w_d} \prod_{i=1}^{n} P(Y_i \mid X_i, w_0, \ldots, w_d),$$

we can consider maximum conditional a posteriori (M(C)AP) estimation:

$$\max_{w_0,\ldots,w_d} \left(\prod_{i=1}^{n} P(Y_i \mid X_i, w_0, \ldots, w_d) \right) P(w_0, \ldots, w_d)$$

where $P(w_0,\ldots,w_d)$ is a prior on the weights. Assuming a standard Gaussian prior $N(0,I)$ for the weight vector, derive the gradient ascent update rules for the weights.

solution

$$P(\omega \mid D) \propto P(\omega)P(D \mid \omega)$$

$$P(\omega) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}\omega^t I^{-1}\omega\right) = \frac{\exp(-\frac{1}{2}\omega^t I^{-1}\omega)}{(2\pi)^{\frac{d}{2}}}$$

$$\omega = \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_d \end{bmatrix}$$

$$P(D \mid \omega) = \prod_{i=1}^{n} P(y_i \mid x_i;\omega)$$

The log of the posterior is then:

$$\hat{\omega} = \arg\max_{\omega} P(\omega \mid D) = \arg\max_{\omega} (\log(P(\omega \mid D)))$$

$$= \arg\max_{\omega} \left(\sum_{i=1}^{n} \log(P(y_i \mid x_i;\omega)) - \frac{d}{2}\omega^\top I^{-1}\omega\right)$$

$$f(\omega) = \sum_{i=1}^{n} \log(P(y_i \mid x_i;\omega)) - \frac{d}{2}\omega^\top I^{-1}\omega$$

To find the weights that maximize this posterior, we use gradient ascent. The gradient of the log-posterior is:

$$\text{Gradient ascent:} \quad \omega_i^{(t+1)} = \omega_i^{(t)} + \eta \nabla f(\omega)$$

$$\frac{\partial f}{\partial \omega_i} = -\omega_i + \sum_{i=1}^{n} \frac{1}{P(y_i \mid x_i;\omega)} \frac{\partial P(y_i \mid x_i;\omega)}{\partial \omega_i}$$

$$\frac{\partial f}{\partial \omega_i} = -\omega_i + \sum_{i=1}^{n} x_i y_i \left(\frac{1}{1+e^{-\omega^\top x_i}}\right) - x_i \frac{e^{-\omega^\top x_i}}{(1+e^{-\omega^\top x_i})^2}$$

$$P(y=1 \mid x_i;\omega) = \frac{1}{1+e^{-\omega^\top x_i}}$$

$$\Rightarrow \frac{\partial P(y=1 \mid x_i;\omega)}{\partial \omega_j} = x_{ij} \frac{e^{-\omega^\top x_i}}{(1+e^{-\omega^\top x_i})^2} = x_{ij} \left(\frac{1}{1+e^{-\omega^\top x_i}}\right)\left(1 - \frac{1}{1+e^{-\omega^\top x_i}}\right)$$

$$P(y=0 \mid x_i;\omega) = \frac{e^{-\omega^t x_i}}{1+e^{-\omega^t x_i}}$$

$$\Rightarrow \frac{\partial P(y=0 \mid x_i;\omega)}{\partial \omega_j} = -x_{ij}\frac{e^{-\omega^t x_i}}{(1+e^{-\omega^t x_i})} + x_{ij}\left(\frac{e^{-\omega^t x_i}e^{-\omega^t x_i}}{(1+e^{-\omega^t x_i})^2}\right)$$

$$\frac{\partial P(y=1 \mid x_i,\omega)}{\partial \omega_j} = x_{ij}P(y=1 \mid x_i,\omega)(1-P(y=1 \mid x_i,\omega))$$

$$\frac{\partial P(y_i=0 \mid x_i,\omega)}{\partial \omega_j} = -x_{ij}P(y_i=1 \mid x_i,\omega)P(y_i=0 \mid x_i,\omega)$$

$$\frac{\partial f}{\partial \omega_j} = -\omega_j + \sum_{i=1}^{n} x_{ij}(y_i - P(y=1 \mid x_i,\omega))$$

$$\omega_j^{(t+1)} = \omega_j^{(t)} + \eta\left(-\omega_j^{(t)} + \sum_{i=1}^{n} x_{ij}(y_i - P(y=1 \mid x_i,\omega))\right)$$

$$\omega_j^{(t+1)} = \omega_j^{(t)}(1-\eta) + \eta\sum_{i=1}^{n} x_{ij}\left(y_i - \frac{1}{1+e^{-(\omega^{(t)})^\top x_i}}\right)$$

Naive Bayes

Consider a Naive Bayes classification problem with three classes and two features. One of these features comes from a Bernoulli distribution and the other comes from a Gaussian distribution. Features are denoted by random vector $X = [X_1, X_2]^\top$ and class is denoted by Y.

Prior distribution is:

$$P(Y=0) = 0.5, \quad P(Y=1) = 0.25, \quad P(Y=2) = 0.25$$

Features distribution is:

$$p_{X_1|Y}(x_1 \mid y=c) = \text{Ber}(x_1;\theta_c), \quad p_{X_2|Y}(x_2 \mid y=c) = \mathcal{N}(x_2;\mu_c,\sigma_c^2)$$

Also assume that:

$$\theta_c = \begin{cases} 0.5 & \text{if } c=0 \\ 0.5 & \text{if } c=1 \\ 0.5 & \text{if } c=2 \end{cases}, \quad \mu_c = \begin{cases} -1 & \text{if } c=0 \\ 0 & \text{if } c=1 \\ 1 & \text{if } c=2 \end{cases}, \quad \sigma_c^2 = \begin{cases} 1 & \text{if } c=0 \\ 1 & \text{if } c=1 \\ 1 & \text{if } c=2 \end{cases}$$

part 1

Find $P(Y \mid X_1=0, X_2=0)$(The answer must be a vector in $\mathbb{R}^3$ where the sum of its elements equal to 1).

solution

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} = \frac{P(x_1 \mid y)P(x_2 \mid y)P(y)}{P(x_1)P(x_2)}$$

$$p(x_1=0) = \sum_{k=0}^{2} P(x_1=0 \mid y=k)P(y=k) = \frac{1}{2}$$

$$p(x_2=0) = \sum_{k=0}^{2} P(x_2=0 \mid y=k)P(y=k) = \frac{1}{4\sqrt{2\pi}}[3e^{-\frac{1}{2}}+1]$$

$$\Rightarrow P(x_1=0)P(x_2=0) = \frac{3e^{-\frac{1}{2}}}{8\sqrt{2\pi}} + \frac{1}{8\sqrt{2\pi}}$$

$$P(y=0 \mid x_1=0, x_2=0) = \frac{P(x_1=0 \mid y=0)P(x_2=0 \mid y=0)P(y=0)}{P(x_1=0)P(x_2=0)} = \frac{2e^{-\frac{1}{2}}}{3e^{-\frac{1}{2}}+1}$$

$$P(y=1 \mid x_1=0, x_2=0) = \frac{P(x_1=0 \mid y=1)P(x_2=0 \mid y=1)P(y=1)}{P(x_1=0)P(x_2=0)} = \frac{1}{3e^{-\frac{1}{2}}+1}$$

$$P(y=2 \mid x_1=0, x_2=0) = \frac{P(x_1=0 \mid y=2)P(x_2=0 \mid y=2)P(y=2)}{P(x_1=0)P(x_2=0)} = \frac{e^{-\frac{1}{2}}}{3e^{-\frac{1}{2}}+1}$$

$$\Rightarrow P(y \mid x_1=0, x_2=0, y=0) = \left(\frac{1}{3e^{-\frac{1}{2}}+1}\right) \begin{bmatrix} 2e^{-\frac{1}{2}} \\ 1 \\ e^{-\frac{1}{2}} \end{bmatrix}$$

part 2

Find $p_{X_1|Y}(x_1 \mid y=1)$.

solution

$$P(y=0 \mid x_1=0) = \int_{-\infty}^{\infty} P(y=0 \mid x_1=0, x_2) P(x_2) dx_2 = \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}(x_2+1)^2}}{2\sqrt{2\pi}} dx_2 = \frac{1}{2}$$

$$P(y=1 \mid x_1=0) = \int_{-\infty}^{\infty} P(y=1 \mid x_1=0, x_2) P(x_2) dx_2 = \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}x_2^2}}{4\sqrt{2\pi}} dx_2 = \frac{1}{4}$$

$$P(y=2 \mid x_1=0) = \int_{-\infty}^{\infty} P(y=2 \mid x_1=0, x_2) P(x_2) dx_2 = \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2}(x_2-1)^2}}{4\sqrt{2\pi}} dx_2 = \frac{1}{4}$$

the result of the solutions comes from integral of a gaussian distribution over the real line.

$$\Rightarrow P(y \mid x_1=0) = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix}$$

part 3

Find $p_{X_2|Y}(x_2 \mid y=0)$.

solution

$$P(y \mid x_2=0) = \frac{P(x_2=0 \mid y)P(y)}{\sum_y P(x_2=0 \mid y)P(y)} = \frac{P(x_2=0 \mid y)P(y)}{\frac{1}{4\sqrt{2\pi}}(2e^{-\frac{1}{2}} + 1 + e^{-\frac{1}{2}})}$$

and the numerator have been calculated in the previous parts so we have:

$$P(y \mid x_2=0) = \frac{1}{1+3e^{-\frac{1}{2}}} \begin{bmatrix} 2e^{-\frac{1}{2}} \\ 1 \\ e^{-\frac{1}{2}} \end{bmatrix}$$

part 4

Justify the pattern that you see in your answers.

solution

as we can see the pattern is that $P_{Y|x_2} = P_{Y|x_1,x_2}$.

the reason is that θ_c is the same for all classes and the features are independent of each other. so the probability of each class given the features is the same as the probability of each class given the other feature and the features.

LDA Vs. LR

Consider the one-dimensional feature X and the two-class response Y. We want to show that classification using linear discriminant analysis is equivalent to using a linear regression model. Specifically, if the sample x_i belongs to the first class we have $Y_i = \frac{n_1}{n}$ and if it belongs to the second class we have $Y_i = \frac{n_2}{n}$. Here, n_1 and n_2 are the number of observations from the first and second classes, and also $n = n_1 + n_2$.

part 1

Using the definition of the discriminant function, show that LDA labels the sample X of the second class if:

$$\frac{\mu_2 - \mu_1}{\sigma^2} X > \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} - \log \frac{n_2}{n_1}$$

And otherwise, it is labeled as the first class.

solution

in below question not that X_i is the feature value for observation i,n is the total number of observations and n_1 and $n_2$are the number of observations in class 1 and class 2.
In LDA, we can define the linear discriminant function as below when μ_1 and μ_2 are the means of class 1 and class 2,σ^2 is the common variance assumed for both classes and $\pi_1 = \left(\frac{n_1}{n}\right)$ and $\pi_2 = \left(\frac{n_2}{n}\right)$ are the prior probabilities of class 1 and class 2.

$$\delta_1(X) = X \cdot \frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \log(\pi_1)$$
$$\delta_2(X) = X \cdot \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \log(\pi_2)$$

if $\delta_1(X) > \delta_2(X)$, then the sample is labeled as class 1 and if $\delta_1(X) < \delta_2(X)$, then the sample is labeled as class 2.so we have:

$$\delta_2(X) > \delta_1(X) = X \cdot \frac{\mu_2}{\sigma^2} - \frac{\mu_2^2}{2\sigma^2} + \log(\pi_2) > X \cdot \frac{\mu_1}{\sigma^2} + \frac{\mu_1^2}{2\sigma^2} - \log(\pi_1)$$

$$\Rightarrow X \cdot \frac{\mu_2 - \mu_1}{\sigma^2} > \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} - \log \frac{\pi_1}{\pi_2} = \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} - \log \frac{n_2}{n_1}$$

$$\Rightarrow X \cdot \frac{\mu_2 - \mu_1}{\sigma^2} > \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} - \log \frac{n_2}{n_1}$$

so if $X \cdot \frac{\mu_2 - \mu_1}{\sigma^2} > \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} - \log \frac{n_2}{n_1}$, the sample is labeled as class 2 and otherwise, it is labeled as class 1.

part 2

Show that the least squares estimate of β_1 in the linear regression model $Y_i = \beta_0 + \beta_1 X_i$ is equal to a multiple (which is only dependent on n) of LDA coefficient for X in part 1.

solution

in linear regression, we can define the least squares estimate of β_1 as below:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y. so we have. based on the condition of part 1 we can define $Y_i = -\frac{n}{n_1}$ for class 1 and $Y_i = \frac{n}{n_2}$ for class 2. so we have:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n_1}(X_{1,i} - \overline{X})\left(-\frac{n}{n_1}\right) + \sum_{i=1}^{n_2}(X_{2,i} - \overline{X})\left(\frac{n}{n_2}\right)}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

if we define $\mu_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i}$ and $\mu_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2,i}$ and $\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ we have:

$$\hat{\beta}_1 = \frac{\left(-\frac{n}{n_1}\right)(n_1\mu_1 - n\overline{X}) + \left(\frac{n}{n_2}\right)(n_2\mu_2 - n\overline{X})}{(n-1)\sigma^2}$$

$$\Rightarrow \hat{\beta}_1 = \frac{n(\mu_2 - \mu_1)}{n-1} \cdot \frac{1}{\sigma^2} = \frac{n}{n-1} \cdot \frac{\mu_2 - \mu_1}{\sigma^2}$$

so we can say that the least squares estimate of β_1 in the linear regression model $Y_i = \beta_0 + \beta_1 X_i$ is equal to a multiple of LDA coefficient for X in part 1.

part 3

Using the previous results, conclude that LDA is equal to comparing the output of linear model $\beta_0 + \beta_1 X$ with a constant.

solution

in part 1 we showed that LDA labels the sample X of the second class if:

$$\frac{\mu_2 - \mu_1}{\sigma^2} X > \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} - \log\frac{n_2}{n_1}$$

so we can say that:

$$\frac{n-1}{n} \cdot \hat{\beta}_1 X > \frac{n}{n-1} \cdot \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} - \log\frac{n_2}{n_1}$$

$$\Rightarrow \hat{\beta}_1 X > \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} - \log\frac{n_2}{n_1}$$

$$\Rightarrow \hat{\beta}_0 + \hat{\beta}_1 X > \hat{\beta}_0 + \frac{\mu_2^2 - \mu_1^2}{2\sigma^2} - \log\frac{n_2}{n_1}$$

wich the right hand of the inequality is constant.so we can say that LDA is equal to comparing the output of linear model $\beta_0 + \beta_1 X$ with a constant.

QDA

Consider samples in the form of $x_i^1 = r\cos(\theta_i)$ and $x_i^2 = r\sin(\theta_i)$. For the first group $r = 3$ and for the second group $r = 5$. We have 16 samples from each group and $\theta_i = i \times \frac{\pi}{8}$ for $1 \le i \le 16$. Find the decision boundary using QDA classifier.

solution

$$\hat{\mu}_1 = \frac{1}{16}\sum_{i=1}^{16} x_1^i = \frac{1}{16}\sum_{i=1}^{16} 3\cos(\theta_i) = 0$$

$$\hat{\mu}_2 = \frac{1}{16}\sum_{i=1}^{16} x_2^i = \frac{1}{16}\sum_{i=1}^{16} 5\sin(\theta_i) = 0$$

$$\hat{\sigma_1}^2 = \frac{1}{16}\sum_{i=1}^{16} (x_1^i - \hat{\mu}_1)^2 = \frac{1}{16}\sum_{i=1}^{16} 9\cos^2(\theta_i) = \frac{9}{2}$$

$$\hat{\sigma_2}^2 = \frac{1}{16}\sum_{i=1}^{16} (x_2^i - \hat{\mu}_2)^2 = \frac{1}{16}\sum_{i=1}^{16} 25\sin^2(\theta_i) = \frac{25}{2}$$

$$\Rightarrow \hat{\mu_1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \hat{\mu_2} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\hat{\Sigma}_1 = \begin{bmatrix} \hat{\sigma_1}^2 & 0 \\ 0 & \hat{\sigma_1}^2 \end{bmatrix} = \begin{bmatrix} \frac{9}{2} & 0 \\ 0 & \frac{9}{2} \end{bmatrix}, \quad \hat{\Sigma}_2 = \begin{bmatrix} \hat{\sigma_2}^2 & 0 \\ 0 & \hat{\sigma_2}^2 \end{bmatrix} = \begin{bmatrix} \frac{25}{2} & 0 \\ 0 & \frac{25}{2} \end{bmatrix}$$

now by this information we can define the decision boundary as $p(y = 0 \mid x; \hat{\theta}) = p(y = 1 \mid x; \hat{\theta})$ and by solving this equation we can find the decision boundary.

$$-\frac{1}{2}(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \Sigma_1^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}) + \log\frac{1}{2\pi + |\Sigma_1|} = -\frac{1}{2}(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \Sigma_2^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}) + \log\frac{1}{2\pi + |\Sigma_2|}$$

$$\Rightarrow -\frac{1}{2}(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \Sigma_1^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \Sigma_2^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}) = \log|\Sigma_1|^{-\frac{1}{2}} - \log|\Sigma_2|^{-\frac{1}{2}}$$

$$\Rightarrow (x_1^2 + x_2^2)\left(\frac{1}{9} - \frac{1}{25}\right) = \log\frac{25}{9}$$

$$\Rightarrow x_1^2 + x_2^2 = 3.79^2$$

so by this decision boundary we can say that the decision boundary is a circle with radius of 3.79 it means that if $x_1^2 + x_2^2 > 3.79^2 \Rightarrow \hat{y} = 2$ and if $x_1^2 + x_2^2 < 3.79^2 \Rightarrow \hat{y} = 1$

Hana

Consider a regression problem with all variables and response having mean zero and standard deviation one. Suppose also that each variable has identical absolute correlation with the response:

$$\frac{1}{N}|\langle x_j, y\rangle| = \lambda_j, \quad j = 1, \ldots, p.$$

Let $\hat{\beta}$ be the least-squares coefficient of y on X, and let $u(\alpha) = \alpha X\hat{\beta}$ for $\alpha \in [0,1]$ be the vector that moves a fraction α toward the least squares fit u. Let RSS be the residual sum-of-squares from the full least squares fit.(a) Show that

$$\frac{1}{N}|\langle x_j, y - u(\alpha)\rangle| = (1-\alpha)\lambda_j, \quad j = 1, \ldots, p,$$

and hence the correlations of each x_j with the residuals remain equal in magnitude as we progress toward u. (b) Show that these correlations are all equal to

$$\lambda(\alpha) = \frac{(1-\alpha)\lambda}{\sqrt{(1-\alpha)^2 + \frac{\alpha(2-\alpha)}{N} \cdot \mathrm{RSS}}}$$

and hence they decrease monotonically to zero.

solution

part a)

$$\frac{1}{N}|\langle x_j, y - u(\alpha)\rangle| = \frac{1}{N}|x_j^T(y - \alpha X\hat{\beta})| = \frac{1}{N}|x_j^T(y - \alpha X(X^TX)^{-1}X^Ty)|$$
$$= \frac{1}{N}|x_j^Ty - \alpha x_j^TX(X^TX)^{-1}X^Ty|$$

by defenition of question we know that $\frac{1}{N}|\langle x_j, y\rangle| = \lambda_j$ so we have:

$$X^T(y - u(\alpha)) = X^Ty - \alpha X^TX(X^TX)^{-1}X^Ty = X^Ty - \alpha X^Ty = (1-\alpha)X^Ty$$

$$\Rightarrow \frac{1}{N}|\langle x_j, y - u(\alpha)\rangle| = \frac{1}{N}|x_j^T(y - u(\alpha))| = \frac{1}{N}|x_j^T(1-\alpha)X^Ty| = (1-\alpha)\lambda_j$$

part b)

Given the regression equation and transformations, we have:

$$\langle \hat{y}, y - \hat{y}\rangle = \hat{y}^Ty - \hat{y}^T\hat{y} = \hat{y}^T(X(X^TX)^{-1}X^Ty) - \hat{y}^T(X(X^TX)^{-1}X^Ty) = 0$$

$$\langle y, y - \hat{y}\rangle = \langle y - \hat{y}, y - \hat{y}\rangle + \langle \hat{y}, y - \hat{y}\rangle$$
$$\Rightarrow \mathrm{RSS} = \langle y - \hat{y}, y - \hat{y}\rangle$$

$$\langle y-u(\alpha), y-u(\alpha)\rangle = \langle y-\alpha\hat{y}, y-\alpha\hat{y}\rangle = \langle (1-\alpha)y+\alpha(y-\hat{y}), (1-\alpha)y+\alpha(y-\hat{y})\rangle$$

$$= (1-\alpha)^2\langle y,y\rangle + \alpha^2\langle y-\hat{y}, y-\hat{y}\rangle + 2\alpha(1-\alpha)\langle y, y-\hat{y}\rangle$$

$$= N(1-\alpha)^2 + 2\alpha(1-\alpha)\langle y, y-\hat{y}\rangle + \alpha^2\mathrm{RSS}$$

$$\Rightarrow \lambda(\alpha) = \frac{(1-\alpha)\lambda}{\sqrt{(1-\alpha)^2 + \frac{\alpha(2-\alpha)}{N}\mathrm{RSS}}}$$

Ridge Regression

part 1

Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim N(0, \tau^2 I)$, and Gaussian sampling model $y \sim N(X\beta, \sigma^2 I)$. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ^2 and σ^2.

solution

$$P(\beta \mid y) \propto P(y \mid \beta) P(\beta)$$

we can define the likelihood and prior as:

$$P(y \mid \beta) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) \right)$$

and the prior as:

$$P(\beta) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} \beta^T \Sigma^{-1} \beta \right)$$

so we have:

$$P(\beta \mid y) \propto \exp\left(-\frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) - \frac{1}{2} \beta^T \Sigma^{-1} \beta \right)$$

taking logaritm of the posterior we have:

$$\log P(\beta \mid y) = -\frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) - \frac{1}{2} \beta^T \Sigma^{-1} \beta$$

so we have:

$$\min_X \left\{ (y - X\beta)^T \Sigma^{-1} (y - X\beta) + \beta^T \Sigma^{-1} \beta \right\}$$

we can define the ridge regression as:

$$\min_\beta \left\{ (y - X\beta)^T \Sigma^{-1} (y - X\beta) + \lambda \beta^T \Sigma^{-1} \beta \right\}$$

so we can relate the regularization parameter λ to the variances τ^2 and σ^2 as:

$$\lambda = \frac{\tau^2}{\sigma^2}$$

part 2

Show that the ridge regression estimates can be obtained by ordinary least squares regression on an augmented data set. We augment the centered matrix X with p additional rows $\sqrt{\lambda} I$ and augment y with p zeroes. By introducing artificial data having response value zero, the fitting procedure is forced to shrink the coefficients toward zero.

solution

we can define the ridge regression as:

$$\min_{\beta} \left\{ (y - X\beta)^T \Sigma^{-1} (y - X\beta) + \lambda \beta^T \Sigma^{-1} \beta \right\}$$

we can augment the matrix X with p additional rows $\sqrt{\lambda} I$ and augment y with p zeroes as:

$$X_{aug} = \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix}$$

and:

$$y_{aug} = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

so we can define the ridge regression as:

$$\min_{\beta} \left\{ (y_{aug} - X_{aug}\beta)^T \Sigma^{-1} (y_{aug} - X_{aug}\beta) \right\}$$

so we can obtain the ridge regression estimates by ordinary least squares regression on an augmented data set.

Estimate

Let $x_1, x_2, \ldots, x_n$ be a sequence of independent random variables from normal distribution with variance σ^2 and mean μ. Answer the following questions (assume you know the amount of variance).

Part 1

Compute the estimator MLE for the mean μ.

solution

$$L(\mu \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

we can take the logaritm of the likelihood as:

$$\log L(\mu \mid x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x_i-\mu)^2}{2\sigma^2}$$

now to find the MLE we can take the derivative of the log likelihood with respect to μ and set it to zero as:

$$\frac{\partial}{\partial \mu} \log L(\mu \mid x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2} = 0$$

so we have:

$$\sum_{i=1}^{n} x_i - n\mu = 0$$

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Part 2

Compute the estimator MAP for the mean μ. Assume that the prior distribution of the mean is from a normal distribution with mean μ and variance β^2.

solution

$$L(\mu \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}(*)$$

the prior is:

$$\mu \sim N(\mu_0, \beta^2)$$

getting log of the $(*)$ we have:

$$log(\text{posterior}) = \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(x_i-\mu)^2}{2\sigma^2} + \log \frac{1}{\sqrt{2\pi\beta^2}} - \frac{(\mu-\mu_0)^2}{2\beta^2}$$

to find the MAP we can take the derivative of the log posterior with respect to μ and set it to zero as:

$$\frac{\partial}{\partial\mu} log(\text{posterior}) = \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} - \frac{\mu-\mu_0}{\beta^2} = 0$$

$$\sum_{i=1}^{n} x_i - n\mu - \frac{\mu-\mu_0}{\beta^2} = 0$$

$$\hat{\mu}_{MAP} = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{1}{\beta^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\beta^2}}$$

Part 3

Check how the results of the first two parts change as the value of N increases towards infinity.

solution

As N increases towards infinity, the MLE and MAP estimators will converge to the true value of the mean μ.