

Classification Exercises

Slides on *Introduction to Statistical Learning*, Chapter 4

Calvin Khor

January 2024

Exercise 1 (paraphrased)

Prove that:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \iff \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}$$

Answer This follows once we show that $f : (0, \infty) \rightarrow (0, 1)$, $f(t) = \frac{t}{1+t}$ has the inverse $g(s) = \frac{s}{1-s}$. Since $f(t) = 1 - \frac{1}{1+t}$, it is monotonic and has an inverse. To finish we simply compute

$$f(t)(1+t) = t \iff f(t) = t(1-f(t)) \iff \frac{f(t)}{1-f(t)} = t \iff g \circ f(t) = t.$$

Exercise 2 (paraphrased)

Prove that in LDA, the class with the largest posterior is the class with the largest discriminant.

Answer Already did this.

Exercise 3 (paraphrased)

Suppose that we have $K \geq 1$ classes and if the observation comes from class k , then $X \sim N(\mu_k, \sigma_k^2)$. Show that the Bayes classifier is not linear, and in fact it is quadratic.

Answer The Bayes classifier is the ideal classifier that chooses the class with the largest conditional probability given the observation, i.e. $\operatorname{argmax}_k \mathbb{P}(Y = k|X)$. We are given $X|Y \sim N(\mu_k, \sigma_k^2)$, so that in the earlier notation, $f_k(x) = (2\pi\sigma_k^2)^{-\frac{1}{2}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$.

Recall $p_k(x) = \frac{\pi_k f_k(x)}{\sum_{\ell} \pi_{\ell} f_{\ell}(x)}$. Since the denominator does not depend on k , and since log is increasing, $p_k(x)$ is maximised iff $\log \pi_k - \frac{(x-\mu_k)^2}{2\sigma_k^2} - \log \sigma_k$ is maximised.

The boundary between two classes k, k' is nonlinear when $\sigma_k \neq \sigma_{k'}$ since it is given by the solutions to

$$\log \pi_k - \frac{(x - \mu_k)^2}{2\sigma_k^2} - \log \sigma_k = \log \pi_{k'} - \frac{(x - \mu_{k'})^2}{2\sigma_{k'}^2} - \log \sigma_{k'}.$$

Exercise 4 (not paraphrased)

When the number of features p is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the *curse of dimensionality*, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigate this *curse*.

- (a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X . We assume that X is uniformly (evenly) distributed on $[0, 1]$. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.6$, we will use observations in the range $[0.55, 0.65]$. On average, what fraction of the available observations will we use to make the prediction?

Exercise 4 (not paraphrased)

When the number of features p is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the **curse of dimensionality**, and it ties into the fact that non-parametric approaches often perform poorly when p is large. We will now investigate this **curse**.

- (a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X . We assume that X is uniformly (evenly) distributed on $[0, 1]$. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.6$, we will use observations in the range $[0.55, 0.65]$. On average, what fraction of the available observations will we use to make the prediction? **Answer** 10%

Exercise 4 (not paraphrased...)

- (b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2 . We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observation's response using only observations that are within 10% of the range of X_1 and within 10% of the range of X_2 closest to that test observation. For instance, in order to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$, we will use observations in the range $[0.55, 0.65]$ for X_1 and in the range $[0.3, 0.4]$ for X_2 . On average, what fraction of the available observations will we use to make the prediction?
- (c) Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

Exercise 4 (not paraphrased...)

- (b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2 . We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observation's response using only observations that are within 10% of the range of X_1 and within 10% of the range of X_2 closest to that test observation. For instance, in order to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$, we will use observations in the range $[0.55, 0.65]$ for X_1 and in the range $[0.3, 0.4]$ for X_2 . On average, what fraction of the available observations will we use to make the prediction? **Answer** 1%.
- (c) Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

Exercise 4 (not paraphrased...)

- (b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2 . We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observation's response using only observations that are within 10% of the range of X_1 and within 10% of the range of X_2 closest to that test observation. For instance, in order to predict the response for a test observation with $X_1 = 0.6$ and $X_2 = 0.35$, we will use observations in the range $[0.55, 0.65]$ for X_1 and in the range $[0.3, 0.4]$ for X_2 . On average, what fraction of the available observations will we use to make the prediction? **Answer** 1%.
- (c) Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?
- Answer** $(10\%)^{100} = \underbrace{0.000\cdots 0}_{99} 1.$

Exercise 4 (not paraphrased.....)

- (d) Using your answers to parts (a)–(c), argue that a drawback of K NN when p is large is that there are very few training observations “near” any given test observation.
- (e) Now suppose that we wish to make a prediction for a test observation by creating a p -dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For $p = 1, 2$, and 100, what is the length of each side of the hypercube? Comment on your answer.

Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When $p = 1$, a hypercube is simply a line segment, when $p = 2$ it is a square, and when $p = 100$ it is a 100-dimensional cube.

Exercise 4 (not paraphrased.....)

- (d) Using your answers to parts (a)–(c), argue that a drawback of K NN when p is large is that there are very few training observations “near” any given test observation.

Answer When p is large, the K nearest neighbors may not be very similar to the test observation, and their outputs may not be representative of the true output.

- (e) Now suppose that we wish to make a prediction for a test observation by creating a p -dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For $p = 1, 2$, and 100, what is the length of each side of the hypercube? Comment on your answer.

Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When $p = 1$, a hypercube is simply a line segment, when $p = 2$ it is a square, and when $p = 100$ it is a 100-dimensional cube.

Exercise 4 (not paraphrased.....💧)

- (d) Using your answers to parts (a)–(c), argue that a drawback of K NN when p is large is that there are very few training observations “near” any given test observation.

Answer When p is large, the K nearest neighbors may not be very similar to the test observation, and their outputs may not be representative of the true output.

- (e) Now suppose that we wish to make a prediction for a test observation by creating a p -dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For $p = 1, 2$, and 100, what is the length of each side of the hypercube? Comment on your answer.

Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When $p = 1$, a hypercube is simply a line segment, when $p = 2$ it is a square, and when $p = 100$ it is a 100-dimensional cube.

Answer The volume of a hypercube with sidelength ℓ is ℓ^p . To equal 10%, we therefore need $\ell = (10\%)^{1/p}$. For $p = 100$, $\ell = 0.977$ which is basically the whole space.

Exercise 5

We now examine the differences between LDA and QDA.

- (a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?
- (b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?
- (c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?
- (d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

Exercise 5

We now examine the differences between LDA and QDA.

- (a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set? **Answer training: QDA, test: LDA**
- (b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?
- (c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?
- (d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

Exercise 5

We now examine the differences between LDA and QDA.

- (a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set? **Answer training: QDA, test: LDA**
- (b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set? **Answer training: QDA, test: QDA**
- (c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?
- (d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

Exercise 5

We now examine the differences between LDA and QDA.

- (a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set? **Answer training: QDA, test: LDA**
- (b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set? **Answer training: QDA, test: QDA**
- (c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?
Answer Improve, as QDA is more flexible so needs more data. Also in applications the true boundary will not be perfectly linear.
- (d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

Exercise 5

We now examine the differences between LDA and QDA.

- (a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set? **Answer** training: QDA, test: LDA
- (b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set? **Answer** training: QDA, test: QDA
- (c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?
Answer Improve, as QDA is more flexible so needs more data. Also in applications the true boundary will not be perfectly linear.
- (d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.
Answer Yes but only if the training set is large enough. If your training set is not large enough then this will not be the case due to overfitting.

Exercise 6

Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

- (a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

- (b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

Exercise 6

Suppose we collect data for a group of students in a statistics class with variables $X_1 =$ hours studied, $X_2 =$ undergrad GPA, and $Y =$ receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

- (a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

Answer

$$\mathbb{P}\left(Y \middle| X = \begin{pmatrix} 1 \\ 40 \\ 3.5 \end{pmatrix}\right) = \sigma\left(\begin{pmatrix} -6 \\ 0.05 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 40 \\ 3.5 \end{pmatrix}\right) = 0.37754066879814546... \approx 38\%.$$

- (b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

Exercise 6

Suppose we collect data for a group of students in a statistics class with variables $X_1 =$ hours studied, $X_2 =$ undergrad GPA, and $Y =$ receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6, \hat{\beta}_1 = 0.05, \hat{\beta}_2 = 1$.

- (a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

Answer

$$\mathbb{P}\left(Y \middle| X = \begin{pmatrix} 1 \\ 40 \\ 3.5 \end{pmatrix}\right) = \sigma\left(\begin{pmatrix} -6 \\ 0.05 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 40 \\ 3.5 \end{pmatrix}\right) = 0.37754066879814546... \approx 38\%.$$

- (b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

Answer

$$\begin{pmatrix} -6 \\ 0.05 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ h \\ 3.5 \end{pmatrix} = \sigma^{-1}\left(\frac{1}{2}\right) \Rightarrow h = 20 \times \left(\sigma^{-1}\left(\frac{1}{2}\right) + 6 - 3.5\right) = 50 \text{ hours.}$$

Exercise 7

Suppose that we wish to predict whether a given stock will issue a dividend this year (“Yes” or “No”) based on X , last year’s percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn’t was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Exercise 7

Suppose that we wish to predict whether a given stock will issue a dividend this year (“Yes” or “No”) based on X , last year’s percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$, while the mean for those that didn’t was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

Answer Given: $X|\text{Yes} \sim N(10, 36)$, $X|\text{No} \sim N(0, 36)$ and $\pi_{\text{Yes}} = 0.8$. Then

$$p_{\text{Yes}} = \frac{\pi_{\text{Yes}} f_{\text{Yes}}(x)}{\sum_{\ell \in \{\text{Yes}, \text{No}\}} \pi_{\ell} f_{\ell}(x)} = 0.7518524532975261... \approx 75.2\%.$$

Exercise 8

Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e. $K = 1$) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

Exercise 8

Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e. $K = 1$) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

Answer Logistic regression.

Exercise 8

Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e. $K = 1$) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

Answer Logistic regression. Recall that 1NN performs perfectly on the train data. So an average error of 18% is really a 36% error rate on the test set, while we are given logistic has a 30% error rate.

Exercise 9

This problem has to do with odds.

- (a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?
- (b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

Exercise 9

This problem has to do with odds.

- (a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?
- (b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

Answer $\omega = \frac{p}{1-p}$ iff $p = \frac{\omega}{1+\omega}$.

Exercise 9

This problem has to do with odds.

- (a) On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default?
- (b) Suppose that an individual has a 16% chance of defaulting on her credit card payment. What are the odds that she will default?

Answer $\omega = \frac{p}{1-p}$ iff $p = \frac{\omega}{1+\omega}$. So

(a) $p = \frac{0.37}{1+0.37} = 0.27007299270072993... \approx 27\%$

(b) $\omega = \frac{0.16}{1-0.16} = \frac{4}{21} = 4 : 21 \text{ odds} \approx 1 : 5 \text{ odds.}$

Exercise 10

Equation (4.32) derived an expression for $\log \frac{\mathbb{P}(Y=k|X=x)}{\mathbb{P}(Y=K|X=x)}$ in the setting where $p > 1$, so that the mean for the k th class, μ_k , is a p dimensional vector, and the shared covariance Σ is a $p \times p$ matrix. However, in the setting with $p = 1$, (4.32) takes a simpler form, since the means μ_1, \dots, μ_K and the variance σ^2 are scalars. In this simpler setting, repeat the calculation in (4.32), and provide expressions for a_k and b_{kj} in terms of $\pi_k, \pi_K, \mu_k, \mu_K$, and σ^2 .

Answer equation (4.32) is

$$\log \frac{\mathbb{P}(Y = k|X = x)}{\mathbb{P}(Y = K|X = x)} = \log \frac{\pi_k}{\pi_K} - \frac{1}{2}(\mu_k + \mu_K)^T \Sigma^{-1}(\mu_k - \mu_K) + x^T \Sigma^{-1}(\mu_k - \mu_K)$$

Since the above formula is valid for all $p \geq 1$, we can just substitute $p = 1$ and simplify to get:

$$\log \frac{\mathbb{P}(Y = k|X = x)}{\mathbb{P}(Y = K|X = x)} = \underbrace{\log \frac{\pi_k}{\pi_K} - \frac{\mu_k^2 - \mu_K^2}{2\sigma^2}}_{a_k} + \underbrace{\frac{(\mu_k - \mu_K)}{\sigma^2} x}_{b_{k1}}.$$

Exercise 11

Work out the detailed forms of a_k , b_{kj} , and b_{kjl} in (4.33). Your answer should involve π_k , π_K , μ_k , μ_K , Σ_k , and Σ_K .

Answer recall $f_k(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left(\frac{-1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k)\right)$. So

$$\begin{aligned} \log \frac{\mathbb{P}(Y = k|X = x)}{\mathbb{P}(Y = K|X = x)} &= \log \frac{\pi_k f_k(x)}{\pi_K f_K(x)} \\ &= \log \frac{\pi_k}{\pi_K} - \frac{1}{2} \log \frac{|\Sigma_k|}{|\Sigma_K|} - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) + \frac{1}{2}(x - \mu_K)^T \Sigma_K^{-1}(x - \mu_K) \\ &= \log \frac{\pi_k}{\pi_K} - \frac{1}{2} \log \frac{|\Sigma_k|}{|\Sigma_K|} - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k + \frac{1}{2} \mu_K^T \Sigma_K^{-1} \mu_K + x^T \Sigma_k^{-1} \mu_k - x^T \Sigma_K^{-1} \mu_K - \frac{1}{2} x^T (\Sigma_k^{-1} - \Sigma_K^{-1}) x \\ &= \log \frac{\pi_k}{\pi_K} - \frac{1}{2} \log \frac{|\Sigma_k|}{|\Sigma_K|} - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k + \frac{1}{2} \mu_K^T \Sigma_K^{-1} \mu_K + x^T (\Sigma_k^{-1} \mu_k - \Sigma_K^{-1} \mu_K) - \frac{1}{2} x^T (\Sigma_k^{-1} - \Sigma_K^{-1}) x \end{aligned}$$

To make it similar to (4.32) we could put $\Sigma_{\pm}^{-1} = \frac{\Sigma_k^{-1} \pm \Sigma_K^{-1}}{2}$ so that $\Sigma_k^{-1} = \Sigma_+^{-1} + \Sigma_-^{-1}$ and $\Sigma_K^{-1} = \Sigma_+^{-1} - \Sigma_-^{-1}$. Then RHS of previous page =

$$\begin{aligned}
&= \log \frac{\pi_k}{\pi_K} - \frac{1}{2} \log \frac{|\Sigma_k|}{|\Sigma_K|} - \frac{1}{2} \mu_k^T \Sigma_k^{-1} \mu_k + \frac{1}{2} \mu_K^T \Sigma_K^{-1} \mu_K + x^T (\Sigma_k^{-1} \mu_k - \Sigma_K^{-1} \mu_K) - \frac{1}{2} x^T (\Sigma_k^{-1} - \Sigma_K^{-1}) x \\
&= \log \frac{\pi_k}{\pi_K} - \frac{1}{2} \log \frac{|\Sigma_k|}{|\Sigma_K|} - \frac{1}{2} (\mu_k + \mu_K)^T \Sigma_+^{-1} (\mu_k - \mu_K) - \frac{1}{2} \mu_k \Sigma_-^{-1} \mu_k - \frac{1}{2} \mu_K \Sigma_-^{-1} \mu_K \\
&\quad + x^T \Sigma_+^{-1} (\mu_k - \mu_K) + x^T \Sigma_-^{-1} (\mu_k + \mu_K) \\
&\quad - x^T \Sigma_-^{-1} x.
\end{aligned}$$

Now when $\Sigma_k = \Sigma_K =: \Sigma$ then $\Sigma_+ = \Sigma$ and $\Sigma_- = 0$, recovering (4.32).

Exercise 12

Suppose that you wish to classify an observation $X \in \mathbb{R}$ into 🍎s and 🍊s. You fit a logistic regression model and find that

$$\hat{\mathbb{P}}(Y = \text{🍊} \mid X = x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}.$$

Your friend fits a logistic regression model to the same data using the softmax formulation in (4.13), and finds that

$$\hat{\mathbb{P}}(Y = \text{🍊} \mid X = x) = \frac{\exp(\hat{\alpha}_{\text{🍊}_0} + \hat{\alpha}_{\text{🍊}_1} x)}{\exp(\hat{\alpha}_{\text{🍊}_0} + \hat{\alpha}_{\text{🍊}_1} x) + \exp(\hat{\alpha}_{\text{🍎}_0} + \hat{\alpha}_{\text{🍎}_1} x)}.$$

- (a) What is the log odds of orange versus apple in your model?
- (b) What is the log odds of orange versus apple in your friend's model?

Exercise 12

Suppose that you wish to classify an observation $X \in \mathbb{R}$ into 🍎s and 🍊s. You fit a logistic regression model and find that

$$\hat{\mathbb{P}}(Y = \text{🍊} \mid X = x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}.$$

Your friend fits a logistic regression model to the same data using the softmax formulation in (4.13), and finds that

$$\hat{\mathbb{P}}(Y = \text{🍊} \mid X = x) = \frac{\exp(\hat{\alpha}_{\text{🍊}_0} + \hat{\alpha}_{\text{🍊}_1} x)}{\exp(\hat{\alpha}_{\text{🍊}_0} + \hat{\alpha}_{\text{🍊}_1} x) + \exp(\hat{\alpha}_{\text{🍎}_0} + \hat{\alpha}_{\text{🍎}_1} x)}.$$

- (a) What is the log odds of orange versus apple in your model?
- (b) What is the log odds of orange versus apple in your friend's model?

Exercise 12

Suppose that you wish to classify an observation $X \in \mathbb{R}$ into 🍎s and 🍊s. You fit a logistic regression model and find that

$$\hat{\mathbb{P}}(Y = \text{🍊} | X = x) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}.$$

Your friend fits a logistic regression model to the same data using the softmax formulation in (4.13), and finds that

$$\hat{\mathbb{P}}(Y = \text{🍊} | X = x) = \frac{\exp((\hat{\alpha}_{\text{🍊}_0} - \hat{\alpha}_{\text{🍎}_0}) + (\hat{\alpha}_{\text{🍊}_1} - \hat{\alpha}_{\text{🍎}_1})x)}{1 + \exp((\hat{\alpha}_{\text{🍊}_0} - \hat{\alpha}_{\text{🍎}_0}) + (\hat{\alpha}_{\text{🍊}_1} - \hat{\alpha}_{\text{🍎}_1})x)}.$$

- (a) What is the log odds of orange versus apple in your model? **Answer** $\hat{\beta}_0 + \hat{\beta}_1 x$
- (b) What is the log odds of orange versus apple in your friend's model?

Answer $(\hat{\alpha}_{\text{🍊}_0} - \hat{\alpha}_{\text{🍎}_0}) + (\hat{\alpha}_{\text{🍊}_1} - \hat{\alpha}_{\text{🍎}_1})x$

Exercise 12 cont.

- (c) Suppose that in your model, $\hat{\beta}_0 = 2$ and $\hat{\beta}_1 = -1$. What are the coefficient estimates in your friend's model? Be as specific as possible.
- (d) *typo fixed* Now suppose that you and your friend fit the same two models on a different data set. This time, your friend gets the coefficient estimates $\hat{\alpha}_{\text{orange}_0} = 1.2$, $\hat{\alpha}_{\text{orange}_1} = -2$, $\hat{\alpha}_{\text{apple}_0} = 3$, $\hat{\alpha}_{\text{apple}_1} = 0.6$. What are the coefficient estimates in your model?
- (e) Finally, suppose you apply both models from (d) to a data set with 2,000 test observations. What fraction of the time do you expect the predicted class labels from your model to agree with those from your friend's model? Explain your answer.

Exercise 12 cont.

- (c) Suppose that in your model, $\hat{\beta}_0 = 2$ and $\hat{\beta}_1 = -1$. What are the coefficient estimates in your friend's model? Be as specific as possible.

Answer By equating the 🍊 log odds we get $\hat{\alpha}_{\text{🍊}0} - \hat{\alpha}_{\text{🍎}0} = 2$, and $\hat{\alpha}_{\text{🍊}1} - \hat{\alpha}_{\text{🍎}1} = -1$. There is no further information that can fix the friend's coefficients: 🍎 log odds are already fixed as $1 - \sigma(t) = \sigma(-t)$.

- (d) *typo fixed* Now suppose that you and your friend fit the same two models on a different data set. This time, your friend gets the coefficient estimates $\hat{\alpha}_{\text{🍊}0} = 1.2$, $\hat{\alpha}_{\text{🍊}1} = -2$, $\hat{\alpha}_{\text{🍎}0} = 3$, $\hat{\alpha}_{\text{🍎}1} = 0.6$. What are the coefficient estimates in your model?
- (e) Finally, suppose you apply both models from (d) to a data set with 2,000 test observations. What fraction of the time do you expect the predicted class labels from your model to agree with those from your friend's model? Explain your answer.

Exercise 12 cont.

- (c) Suppose that in your model, $\hat{\beta}_0 = 2$ and $\hat{\beta}_1 = -1$. What are the coefficient estimates in your friend's model? Be as specific as possible.

Answer By equating the 🍊 log odds we get $\hat{\alpha}_{\text{🍊}0} - \hat{\alpha}_{\text{🍎}0} = 2$, and $\hat{\alpha}_{\text{🍊}1} - \hat{\alpha}_{\text{🍎}1} = -1$. There is no further information that can fix the friend's coefficients: 🍎 log odds are already fixed as $1 - \sigma(t) = \sigma(-t)$.

- (d) *typo fixed* Now suppose that you and your friend fit the same two models on a different data set. This time, your friend gets the coefficient estimates $\hat{\alpha}_{\text{🍊}0} = 1.2$, $\hat{\alpha}_{\text{🍊}1} = -2$, $\hat{\alpha}_{\text{🍎}0} = 3$, $\hat{\alpha}_{\text{🍎}1} = 0.6$. What are the coefficient estimates in your model?

Answer Similarly to the above, $\hat{\beta}_0 = 1.2 - 3 = -1.8$, and $\hat{\beta}_1 = -2 - 0.6 = -2.6$.

- (e) Finally, suppose you apply both models from (d) to a data set with 2,000 test observations. What fraction of the time do you expect the predicted class labels from your model to agree with those from your friend's model? Explain your answer.

Exercise 12 cont.

- (c) Suppose that in your model, $\hat{\beta}_0 = 2$ and $\hat{\beta}_1 = -1$. What are the coefficient estimates in your friend's model? Be as specific as possible.

Answer By equating the 🍊 log odds we get $\hat{\alpha}_{\text{🍊}_0} - \hat{\alpha}_{\text{🍎}_0} = 2$, and $\hat{\alpha}_{\text{🍊}_1} - \hat{\alpha}_{\text{🍎}_1} = -1$. There is no further information that can fix the friend's coefficients: 🍎 log odds are already fixed as $1 - \sigma(t) = \sigma(-t)$.

- (d) *typo fixed* Now suppose that you and your friend fit the same two models on a different data set. This time, your friend gets the coefficient estimates $\hat{\alpha}_{\text{🍊}_0} = 1.2$, $\hat{\alpha}_{\text{🍊}_1} = -2$, $\hat{\alpha}_{\text{🍎}_0} = 3$, $\hat{\alpha}_{\text{🍎}_1} = 0.6$. What are the coefficient estimates in your model?

Answer Similarly to the above, $\hat{\beta}_0 = 1.2 - 3 = -1.8$, and $\hat{\beta}_1 = -2 - 0.6 = -2.6$.

- (e) Finally, suppose you apply both models from (d) to a data set with 2,000 test observations. What fraction of the time do you expect the predicted class labels from your model to agree with those from your friend's model? Explain your answer.

Answer Always. Because the log-odds are exactly equal and they define the decision boundaries.

Extra: ESL Exercise 4.2

Ex. 4.2 Suppose we have features $x \in \mathbb{R}^p$, a two-class response, with class sizes N_1, N_2 , and the target coded as $-N/N_1, N/N_2$.

(a) Show that the LDA rule classifies to class 2 if

$$x^T \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_2 + \hat{\mu}_1)^T \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) - \log(N_2/N_1),$$

and class 1 otherwise.

(b) Consider minimization of the least squares criterion

$$\sum_{i=1}^N (y_i - \beta_0 - x_i^T \beta)^2. \quad (4.55)$$

Show that the solution $\hat{\beta}$ satisfies

$$\left[(N-2)\hat{\Sigma} + N\hat{\Sigma}_B \right] \beta = N(\hat{\mu}_2 - \hat{\mu}_1) \quad (4.56)$$

(after simplification), where $\hat{\Sigma}_B = \frac{N_1 N_2}{N^2}(\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T$.

(c) Hence show that $\hat{\Sigma}_B \beta$ is in the direction $(\hat{\mu}_2 - \hat{\mu}_1)$ and thus

$$\hat{\beta} \propto \hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1). \quad (4.57)$$

Therefore the least-squares regression coefficient is identical to the LDA coefficient, up to a scalar multiple.

(d) Show that this result holds for any (distinct) coding of the two classes.

(e) Find the solution $\hat{\beta}_0$ (up to the same scalar multiple as in (c)), and hence the predicted value $\hat{f}(x) = \hat{\beta}_0 + x^T \hat{\beta}$. Consider the following rule: classify to class 2 if $\hat{f}(x) > 0$ and class 1 otherwise. Show this is not the same as the LDA rule unless the classes have equal numbers of observations.

(Fisher, 1936; Ripley, 1996)

(a) is trivial from knowing the discriminant $\delta_k(x) = \log \pi_k + x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k$ so lets do (b). First we fix notation.

$$y_i \in \left\{ -\frac{N}{N_1}, \frac{N}{N_2} \right\} =: \{c_1, c_2\}, \quad \hat{\mu}_k = \frac{1}{N_k} \sum_{i: y_i = c_k} x_i, \text{ and}$$

$$\hat{\Sigma} = \frac{1}{N-2} \left(\sum_{k: y_k = c_1} (x_k - \hat{\mu}_1)(x_i - \hat{\mu}_1)^T + \sum_{k: y_k = c_2} (x_k - \hat{\mu}_2)(x_i - \hat{\mu}_2)^T \right)$$

starting from the least squares side. The functional to minimise is

$$J(\beta_0, \beta) = \sum_i (y_i - \beta_0 - x_i^T \beta)^2 = \sum_{i: y_i = c_1} (c_1 - \beta_0 - x_i^T \beta)^2 + \sum_{i: y_i = c_2} (c_2 - \beta_0 - x_i^T \beta)^2$$

We split the sum because all terms in the identity use the class split. From $\partial_{\beta_0} J = 0$ we get

$$\begin{aligned} 0 &= \sum_{y_i = c_1} (c_1 - \hat{\beta}_0 - x_i^T \hat{\beta}) + \sum_{y_i = c_2} (c_2 - \hat{\beta}_0 - x_i^T \hat{\beta}) \\ &\iff N \hat{\beta}_0 = N_1 c_1 + N_2 c_2 - (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)^T \hat{\beta} \\ &\iff \hat{\beta}_0 = \underbrace{\frac{N_1 c_1}{N} + \frac{N_2 c_2}{N}}_{=: \Delta_1} - \frac{(N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)^T}{N} \hat{\beta} \end{aligned}$$

for the given class labels c_1, c_2 we have that $\Delta_1 = 0$. From $\partial_\beta J = 0$ we get

$$\begin{aligned}
0 &= \sum_{y_i=c_1} (c_1 - \hat{\beta}_0 - x_i^T \hat{\beta}) x_i^T + \sum_{y_i=c_2} (c_2 - \hat{\beta}_0 - x_i^T \hat{\beta}) x_i^T \\
\iff 0 &= (c_1 - \hat{\beta}_0) N_1 \hat{\mu}_1^T + (c_2 - \hat{\beta}_0) N_2 \hat{\mu}_2^T - \hat{\beta}^T \left(\sum_{y_i=c_1} + \sum_{y_i=c_2} \right) x_i x_i^T \\
\iff 0 &= \underbrace{c_1 N_1 \hat{\mu}_1^T + c_2 N_2 \hat{\mu}_2^T}_{=:\Delta_2^T} - \hat{\beta}_0 (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)^T - \hat{\beta}^T \left(\sum_{y_i=c_1} + \sum_{y_i=c_2} \right) x_i x_i^T
\end{aligned}$$

Taking the transpose,

$$0 = \Delta_2 - \hat{\beta}_0 (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2) - \left(\sum_{y_i=c_1} + \sum_{y_i=c_2} \right) x_i x_i^T \hat{\beta}$$

For the given class labels, $\Delta_2 = N(\hat{\mu}_2 - \hat{\mu}_1)$. Notably Δ_2 is the RHS of the required identity. Plugging in the formula we found for $\hat{\beta}_0$:

$$\left(\left(\sum_{y_i=c_1} + \sum_{y_i=c_2} \right) x_i x_i^T - \frac{1}{N} (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2) (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)^T \right) \beta = \underbrace{\Delta_2 - \Delta_1 (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)}_{=:\Delta_3}$$

we note that

$$\begin{aligned}\hat{\Sigma} &= \frac{1}{N-2} \left(\sum_{k:y_k=c_1} (x_k - \hat{\mu}_1)(x_k - \hat{\mu}_1)^T + \sum_{k:y_k=c_2} (x_k - \hat{\mu}_2)(x_k - \hat{\mu}_2)^T \right) \\ &= \frac{1}{N-2} \left(\sum_k x_k x_k^T + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T - 2N_1 \hat{\mu}_1 \hat{\mu}_1^T - 2N_2 \hat{\mu}_2 \hat{\mu}_2^T \right)\end{aligned}$$

so we can write

$$\sum_k x_k x_k^T = (N-2)\hat{\Sigma} + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T.$$

also,

$$\frac{1}{N} (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2) (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)^T = \frac{N_1^2}{N} \hat{\mu}_1 \hat{\mu}_1^T + \frac{N_2}{N} \hat{\mu}_2 \hat{\mu}_2^T + \frac{N_1 N_2}{N} (\hat{\mu}_1 \hat{\mu}_2^T + \hat{\mu}_2 \hat{\mu}_1^T)$$

so we get out that

$$\begin{aligned}\left((N-2)\hat{\Sigma} + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T - \frac{N_1^2}{N} \hat{\mu}_1 \hat{\mu}_1^T - \frac{N_2}{N} \hat{\mu}_2 \hat{\mu}_2^T - \frac{N_1 N_2}{N} (\hat{\mu}_1 \hat{\mu}_2^T + \hat{\mu}_2 \hat{\mu}_1^T) \right) \beta \\ = \Delta_3\end{aligned}$$

Using that $N_1 + N_2 = N$, we can write $N_1 - \frac{N_1^2}{N} = \frac{N_1(N-N_1)}{N} = \frac{N_1N_2}{N}$, ie. that

$$\begin{aligned}
& N_1\hat{\mu}_1\hat{\mu}_1^T + N_2\hat{\mu}_2\hat{\mu}_2^T - \frac{N_1^2}{N}\hat{\mu}_1\hat{\mu}_1^T - \frac{N_2^2}{N}\hat{\mu}_2\hat{\mu}_2^T - \frac{N_1N_2}{N}(\hat{\mu}_1\hat{\mu}_2^T + \hat{\mu}_2\hat{\mu}_1^T) \\
&= \frac{N_1N_2}{N}(\hat{\mu}_1\hat{\mu}_1^T + \hat{\mu}_2\hat{\mu}_2^T - (\hat{\mu}_1\hat{\mu}_2^T + \hat{\mu}_2\hat{\mu}_1^T)) \\
&= \frac{N_1N_2}{N}(\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T \\
&=: N\hat{\Sigma}_B
\end{aligned}$$

This shows that

$$((N-2)\hat{\Sigma} + N\hat{\Sigma}_B)\hat{\beta} = \Delta_3$$

To finish, we rewrite the expression for Δ_3 :

$$\begin{aligned}
\Delta_3 &= \Delta_2 - \Delta_1(N_1\hat{\mu}_1 + N_2\hat{\mu}_2) \\
&= c_1N_1\hat{\mu}_1 + c_2N_2\hat{\mu}_2 - \left(\frac{N_1}{N}c_1 + \frac{N_2}{N}c_2\right)(N_1\hat{\mu}_1 + N_2\hat{\mu}_2) \\
&= c_1N_1\hat{\mu}_1 + c_2N_2\hat{\mu}_2 - \frac{N_1^2}{N}c_1\hat{\mu}_1 - \frac{N_2^2}{N}c_2\hat{\mu}_2 - \frac{N_1N_2}{N^2}(c_1\hat{\mu}_2 + c_2\hat{\mu}_1) \\
&= \frac{N_1N_2}{N^2}(c_1\hat{\mu}_1 + c_2\hat{\mu}_2 - (c_1\hat{\mu}_2 + c_2\hat{\mu}_1)) \\
&= \frac{N_1N_2}{N^2}(c_1 - c_2)(\hat{\mu}_1 - \hat{\mu}_2)
\end{aligned}$$

We finally arrive at

$$((N-2)\hat{\Sigma} + N\hat{\Sigma}_B)\hat{\beta} = \frac{N_1N_2}{N^2}(c_1 - c_2)(\hat{\mu}_1 - \hat{\mu}_2)$$


(c) is now trivial; this follows from the fact that $ab^Tc = a(b^Tc)$ which is a scalar multiple of the vector a .

(d) was answered by computing in the above with general $c_{1,2}$.

(e) let λ be such that $\hat{\beta} = \lambda \Sigma^{-1} (\hat{\mu}_1 - \hat{\mu}_2)$. We obtain

$$f(x) = \hat{\beta}_0 + x^T \hat{\beta} > 0 \iff \Delta_1 + \lambda \left(x - \frac{N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2}{N} \right)^T \Sigma^{-1} (\hat{\mu}_1 - \hat{\mu}_2) > 0$$

We see that this is equivalent to LDA iff $\frac{\Delta_1}{\lambda} = \log\left(\frac{N_2}{N_1}\right)$. In the original labelling (where $\Delta_1 = 0$) this happens iff $N_1 = N_2$.

- ISLP  python™ Lab videos on Chapter 4
- [Link to labs on github](#)
- [Link to labs in the ISLP !\[\]\(c176e0b06f6c5dd85a4598b214d1ebba_img.jpg\) python™ package documentation](#)
 - NB also includes more information on the datasets, how to use the package etc.