Introduction to Statistical Learning

Author: Robert Hastie et al
Type: #source #textbook
Topics: Statistics Machine Learning

1 - Introduction

- **Statistical learning** refers to tools to understand data; it can be **supervised** or **unsupervised**
- Minimal linear algebra knowledge required for this book
- Notation: n is num. of data points/observations, p is num. of variables, $\mathbf{X}$ is $n \times p$ matrix with x_{ij} representing each observation ($1 \leq i \leq n$, $1 \leq j \leq p$)
    - y_i is ith observation on variable,

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

2 - Statistical Learning

2.1 - What is Statistical Learning?

- Inputs are known as **predictors**, **independent variables**, **features**, **variables**
- If we observe p predictors $X = (X_1, X_2, \ldots, X_p)$ and response Y, there might be a relationship $Y = f(X) + \epsilon$
    - f represents **systematic information** and ϵ is an **error term**
- Statistical learning refers to a set of approaches for estimating f

2.1.1 - Why Estimate f?

- We estimate f for **prediction** or **inference**
    - Prediction: since error ϵ averages to zero, we can predict Y using $\hat{Y} = \hat{f}(X)$
        - $\hat{f}$ is treated as a black-box (do not need to know exact form)
        - Accuracy of $\hat{Y}$ is **reducible error**, ϵ is **irreducible error** (always provides an upper bound for accuracy of prediction of Y)
    - Inference: $\hat{f}$ is not treated as a black box, exact form is needed
        - Important questions:
            - What predictors are associated with response?
            - What is the relationship between response and each predictor?
            - Can the relationship between Y and each predictor be modeled linearly, or is it more complicated?
    - Linear models may be more suitable for simple/interpretable inference, whereas non-linear models may be better for more accurate (but more challenging) prediction

2.1.2 - How Do We Estimate f?

- Goal: apply a statistical learning method to training data to estimate unknown function f (find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$)
- Methods
  - **Parametric methods** consist of a two-step model-based approach
    - 1. Assume f has the form

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

    - 2. Use a procedure to estimate parameters $\beta_0, \beta_1, \ldots, \beta_p$ (a common method is least squares)
    - Disadvantage: model chosen may not match true form of f
      - Might be overly simple, but we could also choose a simpler model leading to overfitting/model learning noise
  - **Non-parametric methods** do not make any assumption about f and aim to estimate an f that is as close to the data points as possible
    - Disadvantage: as no assumptions are made, a very large num. of observations is needed for an accurate estimate

2.1.3 - The Trade-Off Between Prediction Accuracy and Model Interpretability

- A more restrictive method is better for inference because of its interpretability

2.1.4 - Supervised vs. Unsupervised Learning

- **Supervised learning** focuses on inference/prediction tasks and is when we use response variables
- **Unsupervised learning** is when we do not have response variables and need to analyze relationships between variables or observations
  - Ex: **cluster analysis**
- If we have n observations with $m < n$ observations having response variables, we can do **semi-supervised learning** (beyond scope of this book)

2.1.5 - Regression Versus Classification Problems

- Variables can take *quantitative* or *qualitative* values
- **Regression** problems involve a quantitative response, **classification** problems involve a qualitative response (this is not always the case)

2.2. - Assessing Model Accuracy

- *There is no free lunch in statistics*, so there is no method/model dominating all others

2.2.1 - Measuring the Quality of Fit

- Most common measure of fit is **mean-squared error** (MSE) given by

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

  where $\hat{f}(x_i)$ is prediction given by $\hat{f}$ of ith observation
- We are interested in how our method does on the test set/previously unseen data, not the training set
  - We want to use the method minimizing $\mathrm{Ave}(y_0 - \hat{f}(x_0))^2$, where (x_0, y_0) is a previously unseen observation not used to train the learning method
  - Choosing the method that minimizes training MSE does not work

    - Many methods estimate coefficients to minimize training MSE (so test MSE can be much larger)
- **Overfitting** occurs when a given method achieves a small training MSE but a large test MSE

2.2.2 - The Bias-Variance Trade-Off

- Expected test MSE has the following decomposition:

$$E(y_0 - \hat{f}(x_0))^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$$

- We need to select a statistical learning method with low bias and low variance
    - **Variance** refers to the amount $\hat{f}$ changes with changes in the training set
    - **Bias** refers to the error introduced by the method of approximation used
        - Ex: linear regression may be too simple of a method for many problems, so using it may lead to
- The relationship between bias, variance, and the test set MSE is known as the **bias-variance trade-off**
- Cross-validation is a way to test MSE using training data

2.2.3 - The Classification Setting

- Accuracy of $\hat{f}$ in classification is done via **training error rate** $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$ where I is an indicator random variable
- The **test error rate** is given by $\mathrm{Ave}(I(y_0 \neq \hat{y}_0))$ where $\hat{y}_0$ is the predicted class label
- Bayes Classifier
    - Motivation: test error rate is minimized when we choose a classifier assigning each observation to most likely class given predictor values
    - Assigning test observation with predictor x_0 to class j such that $P(Y = j|X = x_0)$ is maximized
    - The **Bayes error rate** is $1 - E(\max_j P(Y = j|X))$
        - Error for a particular $X = x_0$ is $1 - \max_j P(Y = j|X = x_0)$
- K-Nearest Neighbors
    - We do not know the distribution $P(Y = j|X)$ for real data
    - Estimates the conditional distribution of Y given X via $P(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in \mathcal{N}_0} I(y_i = j)$ where K is a given parameter, x_0 is a test observation, and j is a class
        - Using the above estimation, KNN classifies the test observation x_0 to class with largest probability
        - A higher K leads to a low variance but high bias model (choosing an optimal value of K is important as it directly affects model flexibility)

3 - Linear Regression

3.1 - Simple Linear Regression

- **Simple linear regression** is an approach to predict a quantitative respone Y with a single predictor variable X
    - Relationship modeled as $Y \approx \beta_0 + \beta_1 X$
    - Regressing Y **on** or **onto** X
    - β_0, β_1 are **coefficients** or **parameters**; we aim to find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$

3.1.1 - Estimating the Coefficients

- We want to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that $y_i \approx \hat{\beta}_0 + \hat{\beta}_1$ is as "close" as possible to the n data points
    - Defining closeness: a common approach is minimizing the **least squares criterion**
        - If $e_i = y_i - \hat{y}_i$, each e_i is the ith **residual** and the **residual sum of squares** (RSS) is defined as $\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2$, or equivalently as

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

    - Using calculus, we can show the RSS is minimized via

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

      where $\bar{x}$ and $\bar{y}$ are the sample means of x and y, respectively

3.1.2 - Assessing the Accuracy of the Coefficient Estimates

- Note we estimate $Y = f(X) + \epsilon$; this can be written as $Y = \beta_0 + \beta_1 X + \epsilon$ and is known as the **population regression line**
  - Population regression line is the best linear approximation to the true relationship between Y and X
- Information from a sample can be used to estimate characteristics of a large population
  - Ex: multiple different least squares lines generated from random samples of a single population regression line
  - An **unbiased estimator** does not systematically under or overestimate the true parameter
- Accuracy of sample estimators
  - The **standard error** of $\hat{\mu}$ is $\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$
  - If $\sigma^2 = \mathrm{Var}(\epsilon)$ (estimated from the data), the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ are

$$\begin{aligned} \mathrm{SE}(\hat{\beta}_0)^2 &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \\ \mathrm{SE}(\hat{\beta}_1)^2 &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

  - Estimate of σ is the **residual standard error** and given by $\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$
- Confidence intervals
  - Ex: An $\alpha\%$ **confidence interval** is defined in a range of values such that with $\alpha\%$ probability, the interval will contain the true unknown value of the parameter
  - Linear regression 95% CI for $\hat{\beta}_1$ and $\hat{\beta}_2$: $\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$ and $\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$
- Hypothesis testing
  - Most common hypothesis test involves testing **null hypothesis** versus **alternative hypothesis**
    - Null hypothesis: H_0: there is no relationship between X and Y
    - Alternative hypothesis H_a: there is some relationship between X and Y
  - In linear regression, this corresponds to testing $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$
    - How large should $\hat{\beta}_1$ be for us to reject null hypothesis? This also depends on $\mathrm{SE}(\hat{\beta}_1)$
      - To determine this, we compute the t-statistic $t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$
        - Measures number of stdevs $\hat{\beta}_1$ is away from
        - If no relationship between X and Y, we would expect the t-statistic to have a t-distribution with $n-2$ degrees of freedom
  - The probability of observing any number greater than or equal to $|t|$ (assuming $\beta_1 = 0$ is easy in a t-distribution due to its bell shape; this probability is the **p-value**
  - A small p-value indicates it is unlikely to observe a substantial association between X and Y due to chance (small p-value means we reject the null hypothesis)

3.1.3 - Assessing the Accuracy of the Model

- After rejecting null hypothesis that $\hat{\beta}_1 = 0$ in favor of the alternative hypothesis, we want to quantify the extent to which the model fits the data
- Quality of a linear regression fit is assessed using RSE and R^2 statistic
    - Residual standard error (RSE) is equal to $\sqrt{\frac{1}{n-2}\text{RSS}}$
    - R^2 statistic is a proportion (unlike RSE, which is measured in units of Y:

      $$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{RSS}} = 1 - \frac{\text{TSS}}{\text{RSS}}$$

      where $\text{TSS} = \sum(y_i - \bar{y})^2$
        - RSS measures proportion of variability in Y that can be explained by X, so R^2 statistic measures proportion of variability in Y that is not explained by X
    - May also use $\text{Cov}(X, Y)$ (technically $\widehat{\text{Cov}}(X, Y)$ but we omit this for notation)
        - In the simple linear regression setting, it can be shown $R^2 = r^2$

3.2 - Multiple Linear Regression

- We can extend the simple linear regression model for p predictors with

  $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p + \epsilon$$

  where X_j represents the jth predictor and β_j represents the association of X_j and the response
    - β_j is interpreted as the average effect on Y with a one unit increase in X_j (holding all other predictors fixed)

3.2.1 - Estimating the Regression Coefficients

- We aim to estimate $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ to make predictions
    - Parameters are estimated using least squares and minimizing the sum of squared residuals
- A counterintuitive result
    - A variable X_i may have a statistically non-significant positive β_i in a multiple regression while having a significant positive β_1 in a simple linear regression
    - Ex: running a multiple linear regression of shark attacks onto ice cream sales and temperature versus a simple linear regression of shark attacks onto ice cream sales

3.2.2 - Some Important Questions

- Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting Y?
    - We need to ask whether $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ (this is our null hypothesis H_0), whereas the alternative hypothesis is H_a = at least one β_j is non-zero
    - Hypothesis test is performed using F-statistic $F = \frac{(\text{TSS}-\text{RSS})/p}{\text{RSS}/(n-p-1)}$
        - If linear model assumptions are correct, we can show $E(\text{RSS}/(n-p-1)) = \sigma^2$
        - Provided H_0 is true, we can also show $E((\text{RSS} - \text{RSS})/p) = \sigma^2$, so when H_0 is true, the F-statistic should close to 1
    - If n is sufficiently large, an F-statistic closer to 1 may still be sufficient to reject H_0 (on the other hand, a larger F is needed to reject H_0 when n is small)
    - Testing a subset of coefficients: $H_0 = \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_n = 0$
        - We fit a second model that uses all variables except first q
    - Must look at F-statistic even if individual p-values indicate some relationship between Y and X_i
        - F-statistic adjusts for the number of predictors p
            - If we use individual t-statistics and corresponding p-values, about 5% of p-values will be below 0.05 by chance

  - If $p > n$, there are more coefficients β_j to estimate than observations, so we cannot fit a multiple linear regression model using least squares and therefore cannot use F-statistic
- Do all of the predictors help to explain Y, or only a subset?
  - Task of determining which predictors explain Y is known as **variable selection**
  - Approach: try lots of models and judge quality of each
    - Total of 2^p models of all subsets of predictors, so this is impractical unless p is small
  - Other approaches: forward/backward selection, mixed selection
    - Note backward selection cannot be used if $p > n$
- How well does the model fit the data?
  - $R^2 = \mathrm{Cov}(Y, \widehat{Y})^2$ (the fitted linear model maximizes correlation between all possible models)
  - R^2 value always increases (perhaps slightly) when more predictors are added to model
    - Adding a new variable decreases RSS
    - If only a small increase happens with a new predictor, new predictor may not be necessary
  - Graphical summaries of data can
- Given a set of predictor values, what response should we predict, and how accurate is our prediction?
  - There are three types of uncertainty associated with predictions using $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$
    - The least squares plane is only an estimate for the true population regression plane
      - The inaccuracy from coefficient estimates is **reducible error** (we can compute a confidence interval to quantify how close $\hat{Y}$ is to $f(X)$)
    - Assuming a linear model for $f(X)$ can lead to reducible error (**model bias**)
    - Even if $f(X)$ is known, the response cannot be predicted perfectly due to random error ϵ (**irreducible error**)
      - **Prediction intervals** incorporate reducible/irreducible error

3.3 - Other Considerations in the Regression Model

3.3.1 - Qualitative Predictors

- Predictors with only two levels
  - We can use **one-hot encoding** for predictors with two possible values
  - Using such a variable in the regression equation can be done via

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  - Other encoding schemes can also be used, such as (-1, 1), (0, 1) (flipped one-hot encoding), etc.
- Predictors with more than two levels
  - We can use a similar encoding scheme to the two level case with a **baseline**

3.3.2 - Extensions of the Linear Model

- While linear models work well and provide interpretable results, they make assumptions about **additivity** and **linearity** that may not always apply in practice
  - Additivity: association between predictor X_j and Y does not depend on other predictors
  - Linearity: Change in Y associated with a one-unit change in an X_j is constant
- Removing additivity assumption
  - Simple linear regression model may not account for **interaction** effect
  - Solution: add an **interaction term** $\beta_3 X_1 X_2$ in $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ to make the model

$$Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 \epsilon = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$

    - The **hierarchical principle** states that if we include an interaction in a model, we should also include the main effects, even if their associated p-values imply their coefficients are not statistically significant
    - Interactions in qualitative variables are very similar to those in quantitative variables
  - Non-linear relationships
    - **Polynomial regression** can be used to extend linear models

3.3.3 - Potential Problems

- Non-linearity of the response-predictor relationships
  - **Residual plots** can be used to identify patterns in residuals and adjust model (e.g. by adding non-linear terms such as $X^2, \log X, \sqrt{X}$, etc.) to account for non-linearity
- Correlation of error terms
  - Standard errors for regression coefficients are computed assuming error terms ϵ_i are uncorrelated
    - If there is correlation, estimated standard errors will underestimate true standard errors
    - Such correlations occur in context of **time series data** (e.g. if error terms are positively correlated, we may see **tracking** in the residuals where adjacent residuals have similar values)
- Non-constant variance of error terms
  - **Heteroscedasticity** is when errors e_i do not have a constant variance
    - Standard errors/CIs/hypothesis tests for a linear model make this assumption
- Outliers
  - Residual plots can be used to identify outliers
    - Plotting **studentized residuals** (residuals divided by standard error) can be better to decide how large a residual needs to be before a point is considered an outlier
- High-leverage points
  - Points with **high leverage** have an unusual value for x_i
  - The **leverage statistic** can be used to quantify an observation's leverage for a simple linear regression

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

    - Larger leverage statistic implies an observation has high leverage
- Collinearity
  - **Collinearity** refers to when two or more predictor variables are closely related
  - Effects
    - Reduces accuracy of estimation of regression coefficients
    - Causes standard error $\hat{\beta}_j$ to grow, decreasing t-statistic and potentially causing us to fail to reject $H_0 = \beta_j = 0$
      - As a result, **power** of the hypothesis test (defined as probability of correctly detecting non-zero coefficient) is reduced by collinearity
  - Detecting collinearity
    - Between two variables, look for high values in correlation matrix
    - Between multiple values (**multicollinearity**), compute the **variance inflation factor**

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

    where $R^2_{X_j|X_{-j}}$ is R^2 from regression of X_j onto other predictors
      - A VIF exceeding 5 or 10 (arbitrary) indicates significant collinearity
  - Tackling collinearity can be done via combining collinear variables into a single predictor or dropping one or more of collinear variables

3.4 - The Marketing Plan

- Heavily refers to `Advertising` dataset, so omitted
- Tidbits
  - Use prediction intervals when computing individual responses $Y = f(X) + \epsilon$ and confidence intervals when computing average responses $f(X)$ since prediction intervals will account for irreducible error ϵ

3.5 - Comparison of Linear Regression with K-Nearest Neighbors

- Linear regression is a **parametric** approach and makes assumptions about the form of $f(X)$
- **Non-parametric** methods do not make any assumptions about $f(X)$
- **K-nearest neighbors regression** estimates $f(x_0)$ with

$$\hat{f}(x_0)\frac{1}{K}\sum_{x_i \in \mathcal{N}_0} y_i$$

  - Optimal value for k depends on bias-variance tradeoff (lower K gives high variance, low bias)
- Parametric methods outperform non-parametric methods if parametric form is close to true form of X
  - As non-linearity increases, KNN will outperform linear regression
- The **curse of dimensionality** (little observations relative to number of classes) may hinder non-parametric methods

4 - Classification

- Motivation: **categorical** or **quantitative** response variable
- Predicting a qualitative response is referred to as **classifying**

4.1 - An Overview of Classification

- Some examples of why we use classification: disease detection, spam filtering, disease-finding detection genes

4.2 - Why Not Linear Regression?

- Two main reasons
  - 1. Regression cannot accommodate qualitative outputs with more than 2 classes
  - 2. Regression will not provide meaningful estimates of $P(Y|X)$ (even with just 2 classes)

4.3 - Logistic Regression

- Models probability Y belongs to a category (we will do 2 classes)

4.3.1 - The Logistic Model

- We want to model relationship between $p(X) = P(Y = 1|X)$ and X, using 0/1 coding
  - Model used is $p(X) = \beta_0 + \beta_1 X$
  - Problem: probabilities may not be in $[0, 1]$
  - In logistic regression, we use the **logistic function** to create an S-shaped curve

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

    - Quantity $\frac{p(X)}{1-p(X)} = e^{\beta_0 + \beta_1 X}$ is called **odds** and is in $[0, \infty)$
      - $\log \frac{p(X)}{1-p(X)}$ are **log-odds** or **logits**; the linear regression model $p(X)$ is a logit linear in X

- Model fitting using **maximum likelihood**

4.3.2 - Estimating the Regression Coefficients

- Use method of maximum likelihood; choose estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to maximize

$$\prod_{i:y_i=1} p(x_i) \prod_{i':y'_i=0} (1 - p(x_{x'}))$$

- Statistical significance
  - Use a **z-statistic** (analogous to t-statistic), computed via $\frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}$
  - Large abs value of z-statistic indicates evidence against null hypothesis $H_0 : \beta_1 = 0$ or $p(X) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
  - Estimated intercept β_0 is usually not of interest and used to adjust fitted probabilities to proportion of ones in data

4.3.3 - Making Predictions

- Can use 0/1 encoding for case with two qualitative outputs

4.3.4 - Multiple Logistic Regression

- We can generalize two-predictor case as

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

where $(X_1, \ldots, X_p)$ are predictors; this gives

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_P}}{1 + e^{\beta_0+\beta_1 X_p+\cdots\beta_p X_p}}$$

- Confounding variables
  - Results obtained using one predictor may differ from those obtained using multiple predictors (especially when there is correlation among predictors)
  - This is phenomenon known as **confounding**

4.3.5 - Multinomial Linear Regression

- Motivation: extending two-class logistic regression to $K > 2$ classes
- Idea: select a single class (arbitrarily the kth class K) to serve as a **baseline**
  - $P(Y = k|X = x)$ is

$$P(Y = y|X = x) = \frac{e^{\beta_{k_0}+\beta_{k_1}x_1+\cdots\beta_{k_p}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l_0}+\beta_{l_1}x_1+\cdots+\beta_{l_p}x_p}}$$

for $1 \leq k \leq K-1$ and

$$P(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l_0}+\beta_{l_1}x_1+\cdots\beta_{l_p}x_p}}$$

  - For $1 \leq k \leq K-1$,

$$\log\left(\frac{P(Y = k|X = x)}{P(Y = K|X = x)}\right) = \beta_{k0} + \beta_{k_1}x_1 + \cdots + \beta_{k_p}x_p$$

- The choice of baseline is arbitrary and does not affect the output of the model (fitted values, log odds, z-statistics/p-values)

    - Only affects β terms
- We use **softmax coding** as an alternative coding for multiple linear regression
    - As a result, rather than estimating coeffs for $K-1$ classes, we estimate K coefficients

4.4 - Generative Models for Classification

- Motivation: instead of modeling conditional distribution of Y given X, model distribution of each predictor separately given X and use Bayes' theorem to flip these into estimates of $P(Y=k|X=x)$
    - When X is normally distributed, this is similar to logistic regression
    - Reasons why this method may be needed:
        - 1. When there is significant separation between the two classes
        - 2. If predictors X are distributed approximately normally and sample sizes are relatively small, this method may be more accurate than logistic regression
        - 3. Methods can be naturally extended to case with > 2 response classes
- Bayes' theorem
    - If π_k is **prior probability** of randomly chosen observation comes from kth class, $f_(X) \equiv P(X|Y=k)$ is pdf of X (continuous if X is qualitative, discrete if quantitative)
    - **Bayes' theorem** states that the **posterior probability** $P(Y=k|X=x)$ is

$$P(Y=k|X=x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_l f_l(x)}$$

    - Recall Bayes classifier has lowest error rate (if all the terms in the formula are correctly specified)
        - Estimating π_k is easy if we have sufficient sample size
        - Estimating density $f_k(x)$ is more difficult — we discuss three methods to do so

4.4.1 - Linear Discriminant Analysis for $p = 1$

- Assume we have $p = 1$ predictors; to estimate $f_k(x)$, assume it is **normal** or **Gaussian**
    - Density function is pdf

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2(x-\mu_k)^2}\right)$$

    where μ_k and σ_k^2 are mean/variance for kth class
    - Using the above density $f_k(x)$, we obtain

$$p_k(x) = \frac{\pi_k \exp\left(-\frac{1}{2\sigma_k^2(x-\mu_k)^2}\right)}{\sum_{i=l}^{K} \exp\left(-\frac{1}{2\sigma_k^2(x-\mu_k)^2}\right)}$$

    - Log probability is $\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$
    - Bayes decision boundary is point for which $\delta_1(x) = \delta_2(x)$
- **Linear discriminant analysis** (LDA) method approximates Bayes classifier by estimating $\mu_1, \dots, \mu_K$ and $\pi_1, \dots, \pi_K$ (as we cannot compute Bayes classifier unless we know distribution X is sampled from)
    - Estimates used are $\mu_k = \frac{1}{n_k} \sum_{i;y_i=k} x_i$ and $\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i;y_i=k} (x_i - \hat{\mu}_k)^2$
        - Using this, we compute $\hat{\pi}_k = n_k/n$
    - LDA classifier plugs estimates into formula for $\delta_k(x)$ and assigns observation $X = x$ for which $\hat{\delta}_k(x)$ is maximized (this is the **discriminant function**)

4.4.2 - Linear Discriminant Analysis for $p > 1$

- Assume we have $p > 1$ predictors and that each predictor has a 1D normal distribution
  - If a p-dimensional r.v. X has a multivariate Gaussian distribution, we write $X \sim \mathcal{N}(\mu, \mathbf{\Sigma}_k)$, where $\mathbf{\Sigma}_k$ is the $p \times p$ covariance matrix of X
- LDA classifier
  - Assumes $p > 1$ predictors are drawn from a multivariate Gaussian distribution $N(\mu_k, \mathbf{\Sigma})$, where μ_k is a class-specific mean
  - Bayes classifier assigns observation $X = x$ for which $\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mathbf{\Sigma} \mu_k + \log \pi_k$ is largest
  - Bayes decision boundaries represent set of values x such that $\delta_k(x) = \delta_l(x)$
  - Parameters $\mu_1, \dots, \mu_K$, $\pi_1, \dots, \pi_k$ and $\mathbf{\Sigma}$ are estimated in a similar manner to 1D case
- Error rates
  - A **confusion matrix** displays performance of classification algorithm via true/false positive and negative information
  - We can modify the threshold for $p_k(X)$ if we want make less errors of a certain kind (e.g. predicting whether credit card holders will default or not on their statement)
  - The **ROC (receiver operating characteristic) curve** is a popular graphic for simultaneously displaying 2 types of errors for all possible thresholds
  - True/false positive rates are called **sensitivity** and **specificity**
    - If N is the total number of negative values and P (Type I error, 1-specificity) is the total number of positive values, TP/P is the **true positive rate**, FP/N is the **true negative rate** (1-Type II error, power, sensitivity, recall)

4.4.3 - Quadratic Discriminant Analysis

- **Quadratic discriminant analysis** (QDA) assumes (like LDA) observations from each class are drawn from a Gaussian distribution
  - Other assumptions: observation from kth class is of form $\mathcal{N}(\mu_k, \mathbf{\Sigma}_k)$
    - Each class has its own covariance matrix $\mathbf{\Sigma}_k$
  - With these assumptions, Bayes classifier assigns an observation $X = x$ to the class for which $\delta_k = -\frac{1}{2}(x - \mu_k)^T \mathbf{\Sigma}_k (x - \mu_k) - \frac{1}{2} \log |\mathbf{\Sigma}_k| + \log \pi_k$ is greatest
- LDA vs. QDA
  - QDA estimates $Kp(p+1)/2$ parameters whereas LDA estimates $p(p+1)/2$ parameters
    - More parameters leads to higher variance, so QDA is better for large training sets
    - LDA is less flexible and is better for smaller training sets, but if the assumption that K classes have a common covariance matrix is wrong, LDA can have high bias

4.4.4 - Naive Bayes

- Recall in Bayes' theorem, we estimate $p_k(x) = P(Y = k|X = x)$ in terms of $\pi_1, \dots, \pi_K$ and $f_1(x), \dots, f_K(x)$
  - Estimating $\pi_1, \dots, \pi_K$ is easy via training data, but each $f_i(x)$ is a p-dimensional density function with a mean vector and covariance matrix
    - In LDA/QDA, we estimate each mean vector μ_i and 1/K covariance matrices, respectively
  - In Naive Bayes, we assume within the kth class, the p predictors are independent, or $f_k(x) = f_{k_1}(x_1) \cdots f_{k_p}(x_p)$
    - This is a very powerful assumption, as it implies we no longer have to consider the marginal and joint distributions of the p-dimensional density functions
  - This method introduces some bias, but significantly reduces variance
- After making the naive Bayes assumption, we have

$$P(Y = k|X = x) = \frac{\pi_k \cdots f_{k_1}(x_1) \times \cdots \times f_{k_p}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{l_1}(x_1) \times \cdots \times f_{l_p}(x_p)}$$

- Depending on whether X_j is quantitative or qualitative, there are different methods of estimating the one-dimensional density function f_{k_j} using $x_{1_j}, \ldots, x_{n_j}$
    - If X_j quantitative, there are two methods:
        - We can assume $X_j | Y = k \sim \mathcal{N}(\mu_{jk}, \sigma_{jk}^2)$
        - Use a non-parametric estimate for f_{kj}
    - If X_j qualitative, we can count proportion of training observations for the jth predictor corresponding to each class

4.5 - A Comparison of Classification Models

4.5.1 - An Analytical Comparison

- We want to analytically compare LDA, QDA, naive Bayes, logistic regression
    - Setting: we have K classes and want to maximize $P(Y = k | X = x)$
        - Equivalently, we can set K as the baseline class and maximize $\log\left(\frac{P(Y=k|X=x)}{P(Y=K|X=x)}\right)$
    - LDA/QDA assume log odds of posterior probabilities are linear/quadratic in x
    - Naive bayes log odds takes the form of a **generalized additive model**
- Takeaways of computing log odds
    - LDA is a special case of naive Base and naive Bayes is a special case of LDA
    - Neither QDA nor naive Bayes is a special case of the other
    - Logistic regression log odds are also a linear function of the predictors, but chosen to maximize the likelihood function
        - LDA should outperform logistic regression (LR) when normality assumption holds (LR performs better when it does not)
    - KNN
        - Should dominate LDA and LR when decision boundary is very non-linear (assuming n is very large and p is small)
        - KNN requires a lot of observations to perform well
        - QDA may be preferred to KNN when decision boundary is very non-linear but n not large enough or p is not very small
        - Does not tell us which predictors are important

4.5.2 - An Empirical Comparison

- Mostly book-specific, but some takeaways are below
    - No one method always dominates the other
    - Similar to in regression, we can perform classification using transformations of the predictors (e.g. X^2, X^3, X^4, etc.)

4.6 - Generalized Linear Models

- This is also book-specific, but some notes on the models used are below
- Motivation
    - Recall the issue of heteroscedasticity — linear models assume error ϵ has mean zero and constant variance
    - Due to continuous nature of error ϵ, linear model may output continuous values when it should output integers

4.6.2 - Poisson Regression

- Recall the Poisson distribution — it is used to measure **counts** and has mean/variance λ

- Consider a model for the mean $\lambda = E(Y)$ with covariates $X_1, \dots, X_p$

$$\log(\lambda(X_1, \dots, X_p)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  or equivalently $\lambda(X_1, \dots, X_p)) = e^{\beta_0 + \beta_1 X_1 + \beta_p X_p}$ where we want to estimate the β_i parameters (we took the log above so the model is linear in $X_1, \dots, X_p$)
- We use a similar maximum likelihood approach to maximize $l(\beta_0, \dots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!}$
- Differences between Poisson and linear regression
  - Interpretation: an increase in X_j by one unit is associated with an increase in the $E(Y) = \lambda$ by e^{β_j}
  - Mean-variance relationship: better modeled by Poisson regression, since variance is not constant (like linear regression assumes) and mean/variance are equal in Poisson regression
  - Nonnegative fitted values: unlike linear regression, Poisson regression returns nonnegative values

4.6.3 - Generalized Linear Models in Greater Generality

- We have discussed three regression modelling approaches: linear, logistic, Poisson, which share some characteristics:
  - Each approach uses predictors $X_1, \dots, X_p$ to predict a response Y (we assume the distribution of Y based on the approach)
  - Each approach models mean of Y as a function of predictors
- We can transform each model for the mean of Y as a function of its predictors using a **link function**
  - Any regression approach that follows this general recipe is a **generalized linear model**

5 - Resampling Methods

- **Resampling methods** involve repeatedly drawing samples from a training set and refitting a model on each sample to obtain information about the fitted model
  - Common resampling methods include **cross-validation** and the **bootstrap**
  - Resampling is used for **model assessment** and **model selection**

5.1 - Cross-Validation

- Motivation: we may not always have a test set to compute the test error of a statistical method
  - To compute test error, we have several methods:
    - Make an adjustment using training error
    - **Hold out** a subset of the training data and apply the method on the held out data

5.1.1 - The Validation Set Approach

- The validation set approach consists of randomly dividing available set of observations into a **training set** and a **validation set** (aka a **holdout set**)
  - Validation set MSE provides estimate of test error rate
- Drawbacks of validation set approach
  - Validation estimate of test error rate can be highly variable and depends on which observations are included in training set and which are in validation set
  - Only a subset of the observations are included in training set, but statistical models perform better with more training data
    - So, validation set error may overestimate test error rate

5.1.2 - Leave-One-Out Cross Validation

- Similar to validation set approach, but instead of creating two subsets, the model is trained on $n-1$ observations from training set and test MSE is computed using a particular (x_i, y_i)
    - Leads to more variability, but MSE_i is approximately an unbiased estimator of test error
    - Procedure can be repeated with each of $(x_1, y_1), \ldots, (x_n, y_n)$ used to estimate test error
        - LOOCV estimate for test MSE is avg of n test error estimates, e.g. $\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i$
- Advantages
    - Does not overfit test error rate as much as validation set approach because training set is larger
    - No randomness in performing LOOCV, so results are always the same for each (x_i, y_i)
- Optimization
    - We can make the cost of LOOCV the same as that of a single model fit via the formula $\mathrm{CV}_{(n)} = \sum_{i=1}^{n} \left(\frac{(y_i - \hat{y}_i^2)}{1-h_i}\right)$ where h_i is leverage (defined in Chapter 3)
        - This formula does not always hold, in which case the model must be fit n times

5.1.3 - k-Fold Cross Validation

- Approach: divide set of observations into k **folds** of roughly equal size
    - First fold is treated as validation set, method is fit on $k-1$ other folds
    - MSE is computed on held-out fold
    - Procedure is repeated k times, where a different group of observations is treated as validation set
    - k-fold CV estimate is computed via $\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i$
- Remark: LOOCV is a special case of k-fold CV where $k = n$
- Evaluating k-fold CV results
    - If we want to determine how well a given model might perform on independent data, we may look for the model with the best estimate of the MSE
    - Alternatively, if we want the model with the lowest test error, in which we look at the lowest point on the estimated test MSE curve

5.1.4 - Bias-Variance Trade-Off for k-fold Cross-Validation

- k-fold CV often gives more accurate estimates of test error rate than LOOCV
    - LOOCV gives roughly unbiased estimates of test error
        - However, because each model is trained on roughly the same training data, there is high variance (mean of many highly correlated quantities has high variance)
    - k-fold CV leads to an intermediate amount of bias
        - Models are trained on less correlated training data, so there is less variance when compared to LOOCV
    - Validation set approach leads to high bias
- Based on bias-variance tradeoff, one should consider 5/10-fold CV because it has been shown these estimates do not show excessive bias/variance

5.1.5 - Cross-Validation on Classification Problems

- When Y is qualitative, the LOOCV error rate takes the form $\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Err}_i$ (analogous for validation set and k-fold CV error rates)

5.2 - The Bootstrap

- The **bootstrap** is a method used to estimate the uncertainty associated with a given estimator/learning method
    - Ex: estimating standard errors of coeffs from a linear regression model

    - We cannot generate more training data to estimate uncertainty, but we can repeatedly sample from original dataset
- Method
    - Suppose training dataset is Z with n observations
    - Randomly select n observations *with replacement* from dataset to produce a bootstrap dataset Z^{*1} (do this B times to obtain $Z^{*1}, \dots, Z^{*B}$) and corresponding estimates
    - If our estimates are $\hat{\alpha}^{*1}, \dots, \hat{\alpha}^{*B}$, we compute the standard error via

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1} \sum_{r=1}^{B} \left(\hat{\alpha}^{*r} - \frac{1}{B} \sum_{r'=1}^{B} \hat{\alpha}^{*r'} \right)}$$

        - This serves as an estimate of standard error of $\hat{\alpha}$ estimated from original dataset

6 - Linear Model Selection and Regularization

- While we will eventually consider non-linear models, there is still merit to using a linear model $Y = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p + \epsilon$
    - There are some improvements that can be made to a linear model, particularly with least squares
    - Issues with least squares
        - Prediction accuracy: if n is not much larger than p, least squares can have a lot of variability; if $p > n$, there is no unique least squares solution
        - Model interpretability: linear regression sometimes contains irrelevant variables, adding unnecessary complexity in the model
            - A related problem is **feature selection**
- We will discuss **subset selection**, **shrinkage** (**regularization**), and **dimension reduction**

6.1 - Subset Selection

6.1.1 - Best Subset Selection

- Algorithm:
    - 1\. Let $\mathcal{M}_0$ be **null model**, which always predicts the sample mean for each observation
    - 2\. For $k = 1, \dots, p$
        - 2a) Fit least squares regression models with $\binom{p}{k}$ predictors
        - 2b) Pick best model among $\binom{p}{k}$ models (lowest RSS or highest R^2)
    - 3\. Select best model out of all models chosen in 2b using predicted error from a validation set, Bayesian criterion (BIC), or adjusted R^2
- Note training error decreases monotonically with more predictors, so choosing the best model out of all $p + 1$ best models is not trivial
- Can be used for classification problems, but instead of RSS, the **deviance** (negative two times maximized log-likelihood)
- Caveat: very computationally expensive, as we have to evaluate 2^p models

6.1.2 - Stepwise Selection

- Forward stepwise selection
    - Algorithm:
        - 1\. Let $\mathcal{M}_0$ be null model
        - 2\.

            - 2a) Consider all $p-k$ models that augment predictors in $\mathcal{M}_k$ with one additional predictor
            - 2b) Choose the best (in same way as in best subset selection) among these $p-k$ models and call it $\mathcal{M}_{k+1}$
        - 3. Select a single best model among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ in same way as in best subset selection
    - Total models fitted is $1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2$ models (the first model is the null model)
    - Caveat: not guaranteed to find best possible model out of all 2^p models because $\mathcal{M}_{i+1}$ will contain everything in $\mathcal{M}_i$
        - Can be used if $n < p$, but a unique solution will not be reached if $p \geq n$
- Backward stepwise selection
    - Algorithm omitted (similar to forward stepwise selection)
    - Caveat:
- Hybrid approaches between forward/backward stepwise selection are best

6.1.3 - Choosing the Optimal Model

- We need to determine which model from best subset selection/forward/backward selection result in the best model
    - Approaches:
        - 1. Indirectly estimate test error by adjusting training error to account for overfitting
        - 2. Directly estimate test error using a validation set approach or CV
- Adjusting training error
    - C_p is defined as $\frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2)$ where d is the number of predictors in the model
        - Penalty for training error to account for underestimation of test error
    - **Akaike information criterion** (AIC) is defined for models fit with maximum likelihood
        - In the case of a model with Gaussian errors, maximum likelihood and least squares have the same results, so $\text{AIC} = \frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2)$
    - **Bayesian information criterion** (BIC) is defined as $\frac{1}{n}(\text{RSS} + \log(n)d\hat{\sigma}^2)$
        - As $\log(n) > 2$ when $n > 7$, BIC places a higher penalty on larger datasets (higher emphasis on smaller models)
    - The **adjusted** R^2 value is defined as $1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$
        - Large adjusted R^2 indicates a model with small test error
        - Penalizes addition of noise variables in the model via $n-d-1$ term
- Validation/CV
    - Validation/CV may be better for model selection than adjusting training error
        - Less assumptions made, can be used in a wider selection of tasks,
    - The **one-standard error rule** states that the smallest (simples) model within one stdev of the lowest point on the error curve

6.2 - Shrinkage Methods

- Motivation: fit a model containing all p predictors using a technique that **constrains** or **regularizes** coefficient estimates

6.2.1 - Ridge Regression

- **Ridge regression** aims to find estimates $\hat{\beta}^R$ as the values that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{n} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{n} \beta_j^2$$

where $\lambda \geq 0$ i a **tuning parameter** and $\lambda \sum_{j=1}^{n} \beta_j^2$ is a **shrinkage penalty**

  - Note: individual coefficients β_i may increase as λ increases (while ridge coefficient estimate as a whole tend to decrease in aggregate as λ increases)

- Because the estimates $\hat{\beta}_{j,\lambda}$ are not **scale equivariant**, the value of $X_j \hat{\beta}_{j,\lambda}$ depends on the scaling of X_j, as well as potentially other variables $X_1, \ldots, X_p$
  - As a result, it is best to apply ridge regression after **standardizing** the predictors using

$$\bar{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}}$$

- Improvement over least squares
  - When relationship of response to predictors is roughly linear, least squares estimates will have low bias but high variance
  - If $p > n$, the least squares estimates do not have a unique solution, whereas ridge regression can perform well by trading off a small increase in bias for a large decrease in variance
  - Computationally more efficient than best subset selection

6.2.2 - The Lasso

- Motivation: ridge regression always includes all p predictors
- The **lasso** coefficients $\hat{\beta}_\lambda^L$ aim to minimize

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{i=1}^{p} |\beta_j|$$

  - Uses an L1 norm (whereas ridge regress uses an L1 norm)
  - Forces some coefficient estimates to be zero when λ is sufficiently large, so **variable selection** is performed (yields **sparse** models as a result)
- Alternative formulation for ridge and lasso regression
  - We can have a "budget" s such that $\sum_{i=1}^{p} |\beta_j| \leq s$ and $\sum_{i=1}^{p} \beta_j^2 \leq s$ that prevent the coefficient estimates from becoming too small
    - This can be generalized to $\sum_{j=1}^{p} I(\beta_j \neq 0) \leq s$ which represents best subset selection
- Variable selection property
  - Refer to book, but we can essentially represent points containing $\hat{\beta}$ values with the same RSS value on ellipses
    - When an ellipse touches the constraint region (represented by $|\beta_1| + |\beta_2| \leq s$ or $\beta_1^2 + \beta_2^2 \leq s$), lasso/ridge regression stop working
- Comparing lasso/ridge regression
  - Lasso implicitly assumes some predictors will be zero, so ridge will do better in the case where all predictors are related to the response
    - However, as the number of predictors related to the response is not known *a priori*, cross-validation can be used to determine the
  - Shrinkage
    - Ridge regression shrinks every dimension of data by the same proportion
    - Lasso regression shrinks all coefficients towards zero by the same amount
- Bayesian interpretation
  - Assumption: coefficient vector β has some prior distribution $p(\beta)$ where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$
    - If $f(Y|X, \beta)$ is the likelihood of the model, we can multiply the prior distribution by the likelihood to give the **posterior distribution**

$$p(\beta|X,Y) \propto f(Y|X,\beta)p(\beta|X) = f(Y|X,\beta)p(\beta)$$

where we use Bayes theorem to demonstrate proportionality and the fact that X is stationary in the equality

6.2.3 - Selecting the Tuning Parameter

- Cross-validation provides a simple way to choose λ
  - Choose a grid of λ values and compute cross-validation error for each λ
  - Select λ for which cross-validation error is smallest

6.3 - Dimension Reduction Methods

-