Solutions Midterm I for STSCI 3740/5740

Fall, 2024

Course questions (8 points)

Regression
In multiple linear regression, the model is represented in the following matrix form

$$y = X\beta + \varepsilon,$$

where $y \in \mathbb{R}^n$ is the target vector that we wish to predict, $X \in \mathbb{R}^{n \times p}$ is the design matrix and $\beta \in \mathbb{R}^p$ is the parameter to be estimated.

1. **What is the name of the method used to estimate $\beta \in \mathbb{R}^p$.**
Minimization of the Mean Squared Error (MSE)
2. **Prove rigorously, without skipping any step in the computation, that the estimator of $\beta \in \mathbb{R}^p$ given by this method can be written as $\hat{\beta} = (X^T X)^{-1} X^T Y$.**

Consider the linear regression model:

$$Y = X\beta + \varepsilon$$

where:

- Y is an $n \times 1$ vector of observed dependent variable values.
- X is an $n \times p$ matrix of independent variables (with n observations and p predictors).
- β is a $p \times 1$ vector of coefficients to be estimated.
- ε is an $n \times 1$ vector of error terms (assumed to have mean 0 and constant variance).

The OLS estimator $\hat{\beta}$ minimizes the sum of squared residuals, which is given by:

$$S(\beta) = (Y - X\beta)^T (Y - X\beta)$$

Expanding this expression:

$$S(\beta) = Y^T Y - 2Y^T X\beta + \beta^T X^T X\beta$$

To minimize $S(\beta)$, we take the derivative with respect to β and set it equal to zero:

$$\frac{\partial S(\beta)}{\partial \beta} = -2X^T Y + 2X^T X\beta = 0.$$

Assuming $X^T X$ is invertible, the Hessian is positive definite. This ensures that setting the derivative to zero is enough to find the minimizer. After simplifications we find:

$$X^T X\beta = X^T Y$$

Assuming $X^T X$ is invertible, we can solve for β by multiplying both sides by $(X^T X)^{-1}$:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Thus, the OLS estimator is:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

3. **Explain with your own words what is a test hypothesis, what is a p-value.**

Hypothesis testing is a formal method used to decide whether there is enough evidence to support a specific claim or theory about a population. It starts with two opposing statements:
Null Hypothesis H_0: This is the default or baseline statement that assumes no effect or no difference in the population. It represents the status quo or a claim we are trying to disprove. For example, it might state that two groups have the same mean, or that a treatment has no effect.
Alternative Hypothesis H_1: This is the statement that contradicts the null hypothesis. It is what you want to prove. It suggests there is an effect, difference, or relationship. For example, it might claim that the treatment works or that two groups have different means.
The p-value is a measure that helps us assess how compatible our sample data is with the null hypothesis. It is the probability of observing results as extreme as, or more extreme than, what we have in our data, assuming the null hypothesis is true.
Low p-value (typically less than 0.05) suggests that the observed data is unlikely under the null hypothesis, leading us to reject H_0. This indicates that there is evidence in favor of the alternative hypothesis.
High p-value (typically greater than 0.05) suggests that the data is consistent with the null hypothesis, meaning we fail to reject H_0. This doesn't prove the null hypothesis is true, but there isn't enough evidence to say otherwise.

Classification
In a classification problem, we want, given a feature vector X and a qualitative response Y taking values in $\{1, \dots, K\}$ to estimate $\mathbb{P}[Y = k|X = x]$ for every $k \in \{1, \dots, K\}$.
1. **Explain briefly the multiple logistic regression model. How do we estimate the parameters ?**
Sections 4.3.1 and 4.3.2 in ISLR. I expected at least one of the equations 4.1, 4.2, 4.3 or 4.4. Estimation of the parameters is done via maximul likelihood estimation.
2. **Explain briefly the LDA model. How do we estimate the parameters**
Section 4.4.1 or 4.4.2 in ISLR.
3. **What is the Bayes decision boundary ?**
The Bayes decision boundary is the theoretical boundary that separates different classes or groups in a classification problem. It represents the decision surface where the classification probabilities for two or more classes are equal, based on the underlying Bayes optimal classifier. The Bayes classifier assigns a data point x to the class with the highest posterior probability given the observed features x.
Geometrically, the Bayes decision boundary defines the regions in the feature space where one class is more likely than another. It separates areas where one class dominates in terms of the probability of membership.
Theoretically, it provides the optimal decision boundary for a classification problem because it minimizes the probability of misclassification **assuming the true data distribution is known**.
In practice, **we never know the exact underlying distributions that generate the data**, but machine learning algorithms try to approximate the Bayes decision boundary based on the available data.

Problem 1 (6 points)

Assume that we have the regression model

$$Y = f(X) + \varepsilon,$$

where ε is independent of X and $\mathbb{E}(\varepsilon) = 0$, $\mathbb{E}(\varepsilon^2) = \sigma^2$. Suppose that we would like to predict Y for $X = x$. We have learnt that we can define $\hat{Y} = \hat{f}(x)$ as the prediction of Y. We could also consider an alternative prediction $\tilde{Y} = \hat{f}(x) + \varepsilon'$, where ε' has the same distribution as the noise ε and is independent of ε and X.

1. **Compute the mean squared prediction error of** $\tilde{Y}$

$$\begin{aligned}
&\mathbb{E}[(Y - \tilde{Y})^2 | X = x] = \mathbb{E}[(f(X) - \hat{f}(X) + \varepsilon - \varepsilon')^2 | X = x] \\
&= \mathbb{E}[(f(X) - \hat{f}(X))^2 | X = x] + \mathbb{E}[(\varepsilon - \varepsilon')^2 | X = x] + 2\mathbb{E}[(f(X) - \hat{f}(X))(\varepsilon - \varepsilon') | X = x] \\
&= \mathbb{E}[(f(X) - \hat{f}(X))^2 | X = x] + \mathbb{E}[(\varepsilon - \varepsilon')^2] + 2\mathbb{E}[(f(X) - \hat{f}(X)) | X = x]\mathbb{E}[(\varepsilon - \varepsilon') | X = x] \\
&\qquad\qquad (\varepsilon,\ \varepsilon' \text{ and } X \text{ are independent}) \\
&= \mathbb{E}[(f(X) - \hat{f}(X))^2 | X = x] + \mathbb{E}(\varepsilon^2) + \mathbb{E}(\varepsilon'^2) + 2\mathbb{E}(\varepsilon\varepsilon') + 0 \qquad (\mathbb{E}[(\varepsilon - \varepsilon') | X = x] = 0) \\
&= \mathbb{E}[(f(X) - \hat{f}(X))^2 | X = x] + 2\sigma^2.
\end{aligned}$$

2. **Show that** $\mathbb{E}[(Y - \tilde{Y})^2 | X = x] > \mathbb{E}[(Y - \hat{Y})^2 | X = x]$.
We have already shown that

$$\mathbb{E}[(Y - \hat{Y})^2 | X = x] = \mathbb{E}[(f(X) - \hat{f}(X))^2 | X = x] + \sigma^2.$$

So, $\mathbb{E}[(Y - \tilde{Y})^2 | X = x] - \mathbb{E}[(Y - \hat{Y})^2 | X = x] = \sigma^2 > 0$.

Problem 2 (6 points)

Consider a classification problem. Assume that the response variable Y can only take value in $C = \{1, 2, 3\}$. For a fixed x_0, assume that the conditional probability of Y given $X = x_0$ follows

$$P(Y = 1|X = x_0) = 0.2; \quad P(Y = 2|X = x_0) = 0.3; \quad P(Y = 3|X = x_0) = 0.5.$$

1. **Derive the Bayes classifier at $X = x_0$.**
   The Bayes classifier at x_0 outputs the class with the highest probability: $\hat{f}_b(x_0) = 3$.

2. **Derive the corresponding Bayes error rate.**
   The Bayes error rate is $1 - \max\{0.5, 0.3, 0.2\} = 0.5$.

3. **Consider a naive classifier $\hat{f}(x_0)$, called random guessing. That is we use the computer to randomly pick one number from $C = \{1, 2, 3\}$ with equal probability as the label for x_0. Compute the expected test error rate of this classifier. Show that the Bayes error rate is smaller than the expected test error rate for random guessing.**
   This classifier satisfies $P(\hat{f}(x_0) = 1) = 1/3$, $P(\hat{f}(x_0) = 2) = 1/3$, $P(\hat{f}(x_0) = 3) = 1/3$.
   The expected test error rate is

   $$\begin{aligned} &P(y_0 \neq \hat{f}(x_0)) = 1 - P(y_0 = \hat{f}(x_0)) \\ &= 1 - P(y_0 = 1 \mid x_0)P(\hat{f}(x_0) = 1) - P(y_0 = 2 \mid x_0)P(\hat{f}(x_0) = 2) - P(y_0 = 3 \mid x_0)P(\hat{f}(x_0) = 3) \\ &= 1 - 0.5/3 - 0.3/3 - 0.2/3 = 2/3. \end{aligned}$$

   We conclude by saying $0.5 > 2/3$

Answers