Midterm II for STSCI 3740/5740

Nayel Bettache

Fall, 2024

- Read all the instructions carefully.
- This is a closed book exam. You cannot use any electronic devices (e.g., calculator, computer, cell phone) and the lecture notes.
- Please make your writing as clear as possible. If your writing is too hard to recognize, this will lead to reduced score.
- Throughout, we use the notation, models, estimators, etc. introduced in the lecture.
- Please write down your name and NetID on the first page. Good luck!

Time: **75 minutes.**

Name: _____ NetID: _____

Marks: _____

Course questions (12 points)

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 ridge regression ? What is it used for ? What is the definition of the ridge estimator ? (2 points)
Ridge regression is a technique used to address multicollinearity in regression analysis by adding an L2 penalty to the coefficient estimates. It is used primarily to reduce the variance of estimates, leading to more reliable predictions, particularly when predictors are highly correlated or numerous compared to observations. The ridge estimator is defined as $\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y$ where λ is the regularization parameter. This approach balances bias and variance, improving predictive performance in many scenarios.

2. Derive rigorously, without skipping any step in the computation, the closed form of the ridge estimator (4 points)
See review session

3. Explain with your own words what is a test hypothesis, what is a p-value. (2 points)
A statistical hypothesis test is a method of statistical inference used to decide whether the data sufficiently supports a particular hypothesis, usually called the null hypothesis and denoted H_0. A statistical hypothesis test typically involves a calculation of a test statistic. Then a decision is made, either by comparing the test statistic to a critical value or equivalently by evaluating a p-value computed from the test statistic.
A p-value measures the probability of observing the collected data, assuming H_0 is true, with smaller values indicating stronger evidence against H_0. Researchers and practitioners typically compare the p-value to a threshold (often 0.05) to decide whether to reject the null hypothesis, thereby assessing the statistical significance of their finding.

4. What is cross-validation ? What is bootstrap ? When do you use them ? (4 points) Cross-validation is a model validation technique that partitions data into subsets for training and testing to assess model performance on unseen data. The k-fold method, where the data is split into k subsets, is commonly used. Bootstrap is a resampling technique that estimates the distribution of a statistic by sampling with replacement. Use cross-validation to ensure your model generalizes well, especially with limited data, and use bootstrap to quantify uncertainty in estimates.

Problem 1 (10 points, 1 point for each subquestion)

Check the boxes next to the right answers. There can be **one to four** correct answers to each question. One point is assigned to a multiple choice question if and only if all boxes next to the correct answers to this question are checked and no box next to an incorrect answer to this question is checked.

1. Which of the following statements are true

   - ☐ The inflexible models usually yield estimators with low variance and small bias.
   - ☐ The inflexible models usually yield estimators with high variance and small bias.
   - ✓ The inflexible models usually yield estimators with low variance and large bias.
   - ☐ None of the above statements is correct.

2. In which case, we usually prefer the local average method rather than the linear regression

   - ☐ When the number of features p is large and n is small.
   - ✓ When the number of samples n is large and p is small.
   - ☐ When the true regression function is approximately linear.
   - ☐ When the noise variance is large.

3. Assume that the model $Y = f(X) + \varepsilon$ holds, where $\sigma^2 = Var(\varepsilon)$, then

   - ✓ The training mean squared error (MSE) can be smaller than σ^2.
   - ☐ The expected test MSE can be smaller than σ^2.
   - ☐ The expected test MSE is typically smaller when the model is more flexible.
   - ✓ The training MSE is typically smaller when the model is more flexible.

4. Which of the following statements are true

   - ✓ There exists some classifier such that its training error rate is smaller than its test error rate.
   - ✓ The training error rate of the Bayes classifier can be smaller than Bayes error rate.
   - ☐ The expected test error rate of the Bayes classifier can be smaller than Bayes error rate.
   - ☐ The expected test error rate of the KNN classifier can be smaller than Bayes error rate.

5. Assume that the model $Y = f(X) + \varepsilon$ holds, then

   - ✓ The regression function $f(x)$ minimizes the mean squared prediction error.
   - ✓ The mean squared prediction error of $f(x)$ is the irreducible error.
   - ☐ Given an estimator of $f(x)$, its expected test MSE can be smaller than the irreducible error.
   - ✓ Given an estimator of $f(x)$, its training MSE can be smaller than the irreducible error.

6. Which of the following models can be estimated using a linear regression model or least square estimator

   ✓ $Y = \beta_0 + \beta_1 X_1^2 + \varepsilon$
   ☐ $Y = \beta_0 + X_1/(1 + \beta_1 X_1) + \varepsilon$
   ✓ $Y = \beta_0 + \beta_1 X_1/(1 + X_1) + \varepsilon$
   ✓ $Y = \beta_0 + \beta_1 X_1 X_2 + \varepsilon$

7. In a linear regression problem,

   ✓ Forward selection can be used when $n < p$.
   ☐ Backward selection can be used when $n < p$.
   ☐ All subset selection can be used when $n < p$.
   ☐ None of the above three methods can be used when $n < p$.

8. Which of the following statements on the logistic regression are true

   ✓ The logistic regression can estimate the conditional probability $P(Y = 0|X = x)$.
   ☐ When Y has three categories (e.g., $Y \in \{0, 1, 2\}$), we can still use logistic regression.
   ✓ We can estimate the logistic regression by the maximum likelihood method.
   ✓ The logistic regression may not work well when the classes are well separated.

9. Which of the following statements are true

   ☐ The LDA model is more flexible than naive Bayes.
   ✓ When Y has three categories (e.g., $Y \in \{0, 1, 2\}$), we can still use LDA.
   ✓ The training error rate of the QDA model is usually smaller than that from the LDA model.
   ✓ The LDA model (and similarly QDA) may not work when $n < p$.

10. Which of the following statements about KNN are true

    ✓ KNN is a non-parametric classification method.
    ✓ KNN can provide an estimate of the conditional probability $P(Y = 0|X = x)$.
    ☐ KNN classifier is usually more flexible when K becomes large.
    ✓ Compared with LDA, we would prefer KNN, when the Bayes decision boundary is highly nonlinear.

Problem 2 (6 points, 1 point for each question. **1 point will be subtracted if you choose wrongly between two options**.)

Answer the following questions.

(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? Why?**
On the training set, QDA is likely to fit the data better than LDA because QDA has more parameters and greater flexibility. QDA estimates a separate covariance matrix for each class, allowing it to fit the training data more closely. This flexibility often leads to a lower training error for QDA compared to LDA.
On the test set, LDA is expected to perform better. Since the true decision boundary is linear, LDA, which assumes equal covariance matrices across classes (leading to a linear decision boundary), will more closely approximate the Bayes boundary. In contrast, QDA, with its extra flexibility, will estimate a quadratic boundary that can overfit the training data, resulting in higher variance and, consequently, higher test error.

(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? Why?** On the training set, idem.
On the test set, QDA is also expected to perform better if the Bayes decision boundary is non-linear. Since QDA can approximate non-linear boundaries by modeling each class with a separate covariance matrix, it can provide a closer approximation to the true boundary, resulting in lower test error. While QDA is more prone to overfitting than LDA, in this case, its flexibility is beneficial because the underlying Bayes boundary is indeed non-linear. LDA, by contrast, will underfit since its linear boundary cannot capture the non-linear structure of the true decision boundary.

(c) **If the number of covariates p is very large, do we expect LDA or KNN to perform better on the training set? On the test set? Why?** On the training set, KNN is typically expected to perform well because it is a non-parametric, instance-based method that can fit the training data quite closely, especially with a low value of K. In high-dimensional spaces, however, KNN can struggle with the curse of dimensionality: as p increases, the distance between points grows, and neighbors are often farther apart and less informative, potentially increasing the training error. On the other hand, LDA imposes a parametric, linear boundary (assuming normally distributed classes with equal covariances), which tends to be more stable in high-dimensional settings since it relies on estimating a manageable number of parameters. LDA might not fit the training data as precisely as KNN in lower dimensions, but in high dimensions, it is often more robust.
On the test set, LDA is expected to perform better than KNN when p is large. The curse of dimensionality affects KNN more severely in this case because, in high-dimensional spaces, even the closest neighbors tend to be distant and thus less similar to a given test point. This makes KNN's local approach increasingly unreliable, leading to high variance and poor generalization. In contrast, LDA is more stable in high dimensions since it uses a linear decision boundary that avoids fitting noise and doesn't rely on the distances between individual points. By assuming a specific distributional form, LDA can generalize better in high dimensions, resulting in a lower test error than KNN.

Problem 3 (10 points)

Consider a classification problem for (Y, X_1, X_2), where Y is a binary response variable taking values in $C = \{0, 1\}$. Let $h(x)$ be any strictly increasing function such that $h(0) = \frac{1}{2}$. Assume that Y given X_1, X_2 follows from the model

$$P(Y = 1|X_1 = x_1, X_2 = x_2) = h(1 - x_1 - x_2).$$

To make the model reasonable, we require for any real value x, $h(x) \in (0, 1)$ (Note that this property is not useful for solving this question).

(a) **Derive the Bayes classifier at $X_1 = x_1$ and $X_2 = x_2$. (2 points)** The Bayes classifier at $X_1 = x_1$ and $X_2 = x_2$ can be defined as:

$$C^{bayes}(x_1, x_2) = \mathbb{1}(h(1 - x_1 - x_2) \geq 0.5)$$

(b) **Draw the Bayes decision boundary and indicate which region corresponds to the label 1 (You may draw the curve in a two-dimensional plot, with x_1 as the x-axises and x_2 as the y-axises). (2 points)**
The Bayes decision boundary is the locus of points in feature space where the classification probabilities of the two classes are equal. That is, it represents the points where the probability of class 1 equals the probability of class 0.

(c) Consider the following two classifiers. The first classifier is

$$\hat{f}_1(x) = 1, \quad \text{if} \quad x_2 > 0, \quad \text{and} \quad \hat{f}_1(x) = 0, \quad \text{if} \quad x_2 < 0.$$

The second one is

$$\hat{f}_2(x) = 1, \quad \text{if} \quad x_1 < 0, \quad \text{and} \quad \hat{f}_2(x) = 0, \quad \text{if} \quad x_1 > 0.$$

Compute the expected test error rate for the two classifiers $\hat{f}_1(x)$ and $\hat{f}_2(x)$ at $x_1 = x_2 = 1$. Can you tell which classifier has smaller expected test error rate? (3 points)

(d) **Assume that the training data only contain two samples $(Y, X_1, X_2) = (0, 1, 1)$, and $(Y, X_1, X_2) = (1, 1, -1)$. What is the training error rate for the Bayes classifier? What is the training error rate for the classifiers $\hat{f}_1(x)$ and $\hat{f}_2(x)$ in part (c)? (3 points)**