Midterm 1 information and review

– Math 313 Statistics for Data Science

Guangliang Chen

Associate Professor

cheng@hope.edu

Hope College, Fall 2023

Presentation Overview

1. General information
2. Review of the material
3. Practice problems

General information

- **Time and location**: October 20, Friday, in class (50 minutes).
- **Rough range of material to be tested**: Chapters 2, 12, and 3 of the textbook
- **Study resources**: This study guide, lecture slides, and practice problems.
- **Cheat sheet**: One page of notes may be used (do not write anything on the back but save it for midterm 2). No other resources will be allowed during the exam.
- **Electronic device policy**: A scientific calculator (but not cell phones, iPads, laptops, etc.) may also be used, but it is not necessary.

General information (cont'd)

- **Type of questions**: Roughly, there are 3 types of questions:
  - questions that test your basic knowledge of the concepts (for example, multiple choice or short answer questions about the meaning of span or linear independence),
  - questions that are computational and very analogous to lecture examples on 1D PCA and simple linear regression and
  - problems that require you to interpret some plots
- **Important reminder**: Show your work on all exam questions (except the multiple-choiced ones) to receive full credit
  - An incorrect answer without any work will receive zero points
  - An incorrect answer with some correct work will receive partial credit
- **Review session**: We will reserve next Wednesday's class for going through the practice questions and answering any questions you might have.
- **Additional office hours**: 8:30-11:30am, October 20, Friday (I am also available on the preceding day by appointment).

Chapter 2 review

- **General concepts**:
  - What is statistical learning, and supervised vs unsupervised
  - What are the different statistical learning tasks (regression, classification, clustering, and dimensionality reduction)
- **Regression**
  - The mathematical formulation of regression
  - What are the two different regression tasks: prediction vs inference, and what is each about
  - How to estimate the true relationship: parametric vs nonparametric methods
  - Measuring quality of fit in regression: training MSE, test MSE, population MSE, and how they are related
  - The expected test MSE at a new point, as well as the bias-variance tradeoff
- **Classification**
  - What is a classifier, training error rate and test error rate
  - What are Bayes classifiers: posterior probability, Bayes decision boundary and Bayes error rate, kNN classification

Chapter 12 review

- **1D PCA**:
  - Problem setup
  - The SVD-based procedure for finding the maximum-variance direction (but you won't be asked to perform SVD by hand)
  - Interpretation of results: first principal direction/axis (also called first principal component loadings vector), and the first principal component of the data.
  - The absolute/relative amount of scatter of the data explained by the first principal component
- ***k*D PCA**:
  - What are the top k principal directions of the data
  - Interpretation of results: first k principal direction/axis (also called first k principal component loadings vectors), and the first k principal components of the data.
  - The amounts of scatter of the data explained by the first k principal directions individually or cumulatively
  - Ways to choose k when not given: elbow method or 95%
  - Other interpretations of PCA: change of coordinate system, orthogonal best-fit, matrix factorization

- ***k*means clustering**:
  - The *k*means objective function
  - The iterative procedure for attempting to solve the *k*means problem
  - How to initialize *k*means
  - Advantages and disadvantages of *k*means
  - Ways to choose *k* when not given: elbow method
  - Numerical issues: feature scaling and dimension reduction
  - Evaluation of clustering results: confusion matrix and overall clustering accuracy
- **spectral clustering**: not required

Chapter 3 review

- **Simple linear regression**:
  - Problem formulation
  - The least squares criterion and solution
  - Total/regression/residual sums of squares
  - Measure of goodness of fit through R^2
  - Model assumptions and how to check them using different plots
- **Multiple linear regression**:
  - Problem formulation and least squares formulation
  - The normal equation for solving the problem
  - Measure of goodness of fit through adjusted R^2

Chapter 1 Practice

Chapter 1 Problems 1-3 and 5-7 (on pages 63-64 of the book)

Chapter 2 Practice

Consider the following data set consisting of 4 points in $\mathbb{R}^2$: $(0,2),(2,0),(1,3),(3,1)$. Answer the following questions:

1. Sketch the data set on paper. What is its centroid?
2. Remove the centroid from each data point and work with the centered data next.
3. How much scatter does the (centered) data have in total?
4. Which of the two unit-norm directions, $\mathbf{u}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}-1\\1\end{bmatrix}$ (a direction parallel to the long side of the rectangle connecting the four points) and $\mathbf{u}_2 = \frac{1}{\sqrt{10}}\begin{bmatrix}-1\\3\end{bmatrix}$ (a direction parallel to the diagonal line connecting $(1,3)$ and $(2,0)$), can capture more scatter from the data? How do you know?
5. What is the first principal direction of the data? You can use Python to do the computing for you.
6. What is the first principal component of the data, and how much scatter does it capture?

Chapter 3 Practice

Chapter 3 Problems 4 and 7 (on pages 128 and 129).

Consider again the data set on the previous slide.

1. Fit a line by hand calculation to the data according to the least squares criterion. What is the equation of the line?
2. How much scatter do the responses of the data have in total?
3. What are the fitted values and residuals?
4. What fraction of the scatter in the response is explained by the least squares line?
5. What is the prediction, according to the fitted model, at $x_0 = 2.5$? Is it valid to use the model to make a prediction at $x_0 = 5$?