MATH 373: Introduction to Machine Learning

Homework 0

This homework provides a valuable review of key concepts needed for this course. We should work out any issues you may have on these problems.

1. Let X be a continuous random variable with density f_X. Use the CDF method to find the density of $Y = |X|$ in terms of f_X.

2. (Ross) Let X be a random variable taking values in the finite interval $[0, c]$. You may assume that X is discrete, though this is not necessary for this problem.

   (a) Show that $\mathbb{E}X \leq c$ and $\mathbb{E}X^2 \leq c\,\mathbb{E}X$.

   (b) Recall that $\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$. Use the inequalities above to show that

   $$\mathrm{Var}(X) \;\leq\; c^2[u(1-u)] \quad \text{where} \quad u \;=\; \frac{\mathbb{E}X}{c} \in [0,1].$$

   (c) Use the result of part (b) to show that $\mathrm{Var}(X) \leq c^2/4$.

   (d) Use the result in (c) to bound the variance of a random variable X taking values in an interval $[a, b]$ with $-\infty < a < b < \infty$.

3. Recall that the covariance of random variables X, and Y is defined by $\mathrm{Cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y)$. Establish the following identities.

   (a) $\mathrm{Cov}(X) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y)$

   (b) If a, b, c, d are constants, then $\mathrm{Cov}(aX + b, cY + d) = ac\ \mathrm{Cov}(X, Y)$

   (c) If $X_1, \ldots, X_n$ are r.v., then $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{1 \leq i < j \leq n} \mathrm{Cov}(X_i, X_j)$

4. Recall that the variance of a random variable X is defined by $\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$. Establish the following.

   (a) $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$

   (b) If a, b are constants, then $\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$

   (c) $\mathbb{E}X^2 \geq (\mathbb{E}X)^2$.

5. The empirical cumulative distribution function (CDF) of a sample $x = x_1, \ldots, x_m$ is defined by

$$F_x(t) \;=\; m^{-1} \sum_{i=1}^{m} \mathbb{I}(x_i \leq t)$$

The sum in the definition counts the number of data points that are less than or equal to t. Thus $F_x(t)$ is the fraction of data points that are less then or equal to t.

Suppose that x has four points: -3, -1, -1, and 5.

a) Find the following values of the empirical CDF by using the formula above: $F_x(-4)$, $F_x(0)$, $F_x(-1)$, $F_x(6)$

b) Sketch the empirical CDF for this data set as a function of t.

c) For what values of t is $F_x(t) = 0$?

d) For what values of t is $F_x(t) = 1$?

6. Let $\mathbf{A}$ and $\mathbf{B}$ be invertible $n \times n$ matrices.

a) Argue that the product $\mathbf{A}\,\mathbf{B}$ is invertible.

b) Argue that $(\mathbf{A}\,\mathbf{B})^{-1} = \mathbf{B}^{-1}\,\mathbf{A}^{-1}$.

7. Let $\mathbf{A}$ be an $n \times p$ matrix. Consider the matrix product $\mathbf{B} = \mathbf{A}^t\,\mathbf{A}$

a) What are the dimensions of $\mathbf{B}$?

b) Show that $\mathbf{B}$ is symmetric.

c) Show that $\mathbf{B}$ is non-negative definite.

8. Suppose that a scientist at AT&T hires you to analyze a dataset that contains the following information about 1500 of the company's customers:

   - Gender
   - Age
   - State residence
   - Average monthly telephone minutes used over the last year
   - Average monthly data usage (in MB) used over the last year
   - Whether or not the customer renewed their contract this year

   a) The scientist first wants to get a general idea of any potential patterns in his data. In particular, he wants to visualize these trends so that he can explain them to the rest of his team at AT&T. List 2 or 3 potential trends that seem worth exploring in this data set. Explain what kinds of analyses you would conduct to explore the data for statistical trends.

   b) The scientist reports back to you that his team would like to understand what demographic features of the customers (gender, age, residence) explains the most variation in the average data usage of each customer. How might you go about answering this question?

   c) The scientist has another data set like this one on 200 additional customers; however, these 200 customers have not yet decided whether or not they will renew their contract. He wants to predict their renewal so that AT&T can reach out to the customers who will likely not renew. Explain how you would go about helping the scientist in this prediction problem. Explain how, if applicable, cross validation can help with your analysis.

9. Consider the traditional multiple linear regression model of response $Y = (y_1, \ldots, y_n)^T$ on data $X \in \mathbb{R}^{n \times p}$:

$$Y = X\beta + \epsilon$$

   a) What possible assumptions can you make on the error terms $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$?

   b) The least squares estimates of $\hat{\beta}$ are given by:

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

i) Do the least squares estimates above depend on any of the assumptions we can make about ϵ?

ii) Using the least squares estimates and an appropriate assumption on ϵ, show that

$$\hat{\beta} \sim N_p(\beta, \sigma^2(X^T X)^{-1})$$

It follows that making an appropriate assumption on ϵ allows us to make inference about our model. We will revisit this with other regression models in the future.