ECE 3502 / CS 3501: Foundations of Data Analysis

Midterm Exam Guide

Format

- Closed book, no notes, no electronic devices, etc.
- Only formulas you need to remember are **in this study guide**!
- Two questions about probability (Section I below). Will require you to use formulas. You won't have to do the arithmetic to simplify answers, just plug in the numbers in the correct formula.
- Several true/false and multiple choice questions.
- 2-3 "in depth" questions about concepts/understanding, requiring written answers (see example problems below).

Topics Covered

I Probability

(a) Basic Probability Rules

- Equally likely outcomes: $P(A) = \frac{|A|}{|\Omega|}$
- Inclusion-Exclusion Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Complement Rule: $P(A^c) = 1 - P(A)$
- Difference Rule: $P(A - B) = P(A) - P(A \cap B)$

(b) Conditional Probability

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

(c) Multiplication Rule

$$P(A \cap B) = P(A \mid B)P(B)$$

(d) Total Probability Rule

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

(e) Bayes' Rule

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

II Estimation

(a) Maximum Likelihood Estimation
- maximizes likelihood of the data x with respect to the parameter θ: $p(x; \theta)$
- parameter θ is an unknown constant we want to learn from data

(b) Bayesian Estimation
    - computes posterior of the parameter $p(\theta \mid x)$
    - unlike MLE, parameter θ is a random variable
    - know about the Beta distribution as the posterior for the rate θ of a binary variable

III Classification

(a) Using conditional probability $P(C \mid X)$. What does this probability tell us?

(b) What is the likelihood, prior, evidence, posterior?

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}$$

(c) Naïve Bayes
    - what is the "naïve" assumption?
    - how does training work?

(d) Training vs. testing
    - what is the difference? why is it important to separate training/testing data?

IV Hypothesis Testing of two binary variables

(a) Contingency tables
    - you should be able to fill one out given a problem setup and count data

(b) Null hypothesis
    - be able to define it, explain the logic behind it

(c) Fisher Exact Test
    - be able to describe how the p value is computed

V Kmeans clustering and KNN classifier

(a) What is the task of Kmeans and KNN?

(b) How do you assign data points to different groups by these two methods?

(c) What is the difference between Kmeans and KNN?

(d) What is the decision rule (learned from class) for KNN?

VI Linear Regression

(a) Understand the formula (and picture): $y_i = \alpha + \beta x_i + \epsilon_i$

(b) What is least squares fitting? How is it related to maximum likelihood?

(c) How is least squares fitting of the slope equivalent to projection?

(d) Relationship between slope and correlation

(e) R^2 statistic. What is it, and what does it tell you about a regression fit?

Example Problems

1. It is estimated that 6 out of every 1,000 people have autism spectrum disorder, i.e., there is a 0.6% chance of being born with the disorder. Of people with autism spectrum disorder, 80% are male. Let A be the event that a person has autism spectrum disorder, and M be the event that a person is a male. Also, assume in this problem that it is equally likely to be born male or female, that is, $P(M) = P(M^c) = 0.5$.

    (a) What does $P(M \mid A)$ mean in English? What is its value?

    (b) What does $P(A \mid M)$ mean in English? What is its value?

    (c) What is the probability of being a female with autism spectrum disorder? (First write down the probability expression in terms of A and M, and then compute.)

2. You have a sock drawer with 4 red socks and 2 blue socks. If you randomly pull two socks out of the drawer, what is the probability that they match?

3. You are given images of cats and dogs and want to train a naïve Bayes classifier to automatically identify them. The images are represented as d-dimensional random variables: $x = (x^1, x^2, \ldots, x^d)$. (note: the superscript is the dimension). Each data point also has a binary class label c (represented as cat=0 / dog=1).

    (a) What assumption does naïve Bayes make about the likelihood $p(x^1, x^2, \ldots, x^d \mid c)$?

    (b) When does this assumption not work well?

    (c) Say you model the likelihood for each dimension as a normal distribution, $p(x^i \mid c; \mu_i, \sigma_i^2)$, and you specify a prior for the classes, $p(c)$. How does the training of this model work? (In other words, what must be computed to fit this model to your data?)

4. You are analyzing data for a clinical trial of a drug that is meant to shorten the recovery time for the flu. You have two binary variables: whether the patient is given the drug or a placebo (fake drug), and whether the patient recovers faster than normal (yes/no). Here are the results:

Fast?	YES	NO
Placebo	20	20
Drug	30	10

    (a) If you want to do a Fisher exact test of the hypothesis that the drug is effective, what is your null hypothesis?

    (b) Which of the following probabilities would you compute as the p value?

        i. The probability of the null hypothesis given this data or possibly a stronger relationship between drug and recovery.

        ii. The probability of the null hypothesis given exactly this data.

        iii. The probability under the null hypothesis of getting exactly this data.

        iv. The probability under the null hypothesis of getting this data or possibly a stronger relationship between drug and recovery

    (c) Just by looking at the contingency table, would you guess that the hypothesis that the drug is effective is true or not?