18-799 M/SM: Special Topics in Signal Processing: Advanced Machine Learning

Homework 1

DUE February 4, 2015 at noon Pacific

- Follow the homework policy in the syllabus, in particular, do not Google for answers!
- For programming problems, use your favorite general programming language such as MATLAB, Julia, Python, C/C++, Java, R, etc. Please include source code in your submissions.
- You may NOT use specialized machine learning packages to solve these problems unless otherwise specified.
- You must submit your homework using Blackboard before the deadline. There will be a penalty for late submission.
- You may discuss the problems with fellow students, but you must complete your own homework.

1 [15 pts] Probability

1. [4 pts] In a class of 20 students, 10 play soccer, 8 play basketball and 6 don't play any sports. If 5 students were chosen at random, what is the probability that all chosen students play exactly one sport?
2. [11 pts] Comparing dice.

   In this problem, we consider three dice with unusual numbering. The numbers on the dice are:

Die 1	3, 5, 6, 9, 15, 18
Die 2	2, 4, 8, 13, 14, 16
Die 3	1, 7, 10, 11, 12, 17

   All of the dice are unbiased; that is all 6 outcomes are equally likely. A game is played with these dice. Each player has a different die. They each roll their die, and whoever has the higher number wins.

   You may write a program to solve this problem, but you MUST explain your approach. If you choose to write code, you must include it with your submission.

(a) [3 pts] If die 1 and 2 are rolled, what is the probability that die 1 beats die 2?

(b) [3 pts] If die 2 and 3 are rolled, what is the probability that die 2 beats die 3?

(c) [3 pts] If die 3 and 1 are rolled, what is the probability that die 3 beats die 1?

(d) [2 pts] Suppose now that we have a game between 2 players, but we let them pick the die they want. The players choose their die one at a time (i.e, the player going second must choose from the remaining dice). From the results above, is it more advantageous to pick first or second? Which die should be chosen to yield the highest probability of winning?

2 [30 pts] Statistics

1. [15 pts] Let us attempt to estimate the probability (θ) of heads of a coin using both Bayesian and frequentist approach. Our prior belief about θ is that it follows a beta distribution with parameters α and β (i.e Beta(α, β)). We toss the coin n times and denote the observations as $X_1, X_2, \cdots, X_n$.

   (a) What is the posterior distribution $\theta | X_1, X_2, \cdots, X_n$?

   (b) Find the maximum a posteriori estimate of θ (denote it by θ_{MAP}). (This is Bayesian approach)

   (c) Given the observations, what is the maximum likelihood estimate of θ(denote it by θ_{MLE})? (This is frequentist approach)

   Hint: Think about the distribution of $X_1 | \theta$.

2. [15 pts] Now let us compare these two estimators by a simulated experiment. We pick a coin with probability of heads as 0.85.

   (a) Use any (binary) random number generator to simulate a coin toss. Using the parameters of the beta distribution as $\alpha = 3, \beta = 3$ compare θ_{MAP} and θ_{MLE} for 10 observations.

   (b) Repeat the experiment for 100, 1000, 10000 observations and compare the two estimators.

   (c) If we choose $\alpha = 1$ and $\beta = 1$, which estimator performs better? Use 1000 observations here.

   (d) What can we conclude from the previous results?

3 [20 pts] Bayesian Networks

1. [10 pts] Based on Figure 1, state whether the following statements are TRUE or FALSE. Provide brief explanation.

   (a) $P(D, L) = P(D)P(L)$

   (b) $P(A, I) = P(A)P(I)$

(c) $P(A, I|G) = P(A|G)P(I|G)$

(d) $P(C, J|A, G) = P(C|A, G)P(J|A, G)$

(e) $P(E, H|M) = P(E|M)P(H|M)$

2. [10 pts] From Figure 1, what are the blankets of E, F, K, H ?

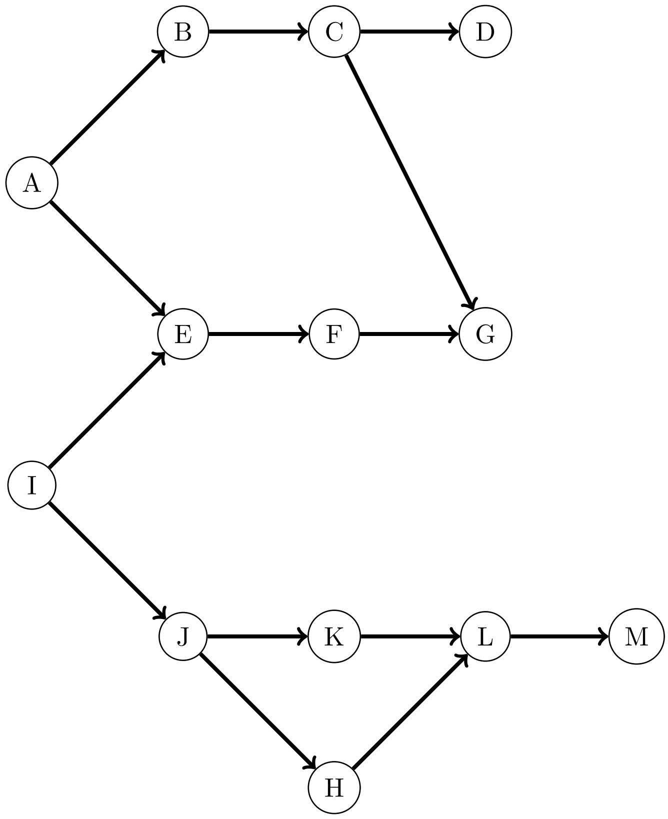

Figure 1: A Bayesian network

4 [35 pts] Classification

This problem involves coding. We will use the Car Evaluation Dataset (http://archive.ics.uci.edu/ml/datasets
from the UCI Machine Learning Repository.

(a) [2 pts] Download the data file and have a look at it. (https://archive.ics.uci.edu/ml/machine-learning-databases/car/car.data). Also take a look at the description of the data in the names file. (https://archive.ics.uci.edu/ml/machine-learning-databases/car/car.names) To pre-process the data, remap the attribute labels into integer values as follows:

Feature	Original Label	Remapped Label
buying	low, med, high, vhigh	1, 2, 3, 4
maint	low, med, high, vhigh	1, 2, 3, 4
doors	2, 3, 4, 5more	1, 2, 3, 4
persons	2, 4, more	1, 2, 3
lug_boot	small, med, big	1, 2, 3
safety	low, med, high	1, 2, 3

For this problem, we're interested in classifying the cars into acceptable vs. unacceptable, so let's remap the class labels as follows:

	Original Label	Remapped Label
class	unacc, acc, good, vgood	-1, 1, 1, 1

How many examples are in class -1? How many are in class 1?

(b) [15 pts] From all these labeled examples (classified as -1 or 1) use randomly sampled 90% examples as training data and the remaining as test data. Implement a Naive Bayes classifier and train it using this training data and compute the percent accuracy on the test data. Repeat this process (sampling, training, testing) 10 times and report the mean accuracy. How confident are you in your estimate of the mean accuracy? Use the standard error of the mean to support your answer.

(c) [15 pts] Now implement a Logistic Regression classifier for this problem. Include the bias term. Again repeat the sampling, training and testing process 10 times as above and report the mean accuracy. How confident are you in your estimate of the mean accuracy? Use the standard error of the mean to support your answer.

Use the gradient descent (GD) algorithm with a learning rate $\alpha = 1 \times 10^{-2}$ to train this classifier. Run GD until the change in the parameter vector goes below $\epsilon = 1 \times 10^{-2}$, i.e., convergence rule: $\|\theta_t - \theta_{t-1}\|_2 < \epsilon$.

(d) [3 pts] Which classifier performs better on this data set? What do you think could be the reason?