University of Washington Bothell
Computing & Software Systems

Course title: Introduction to Machine Learning
Course number: CSS 490 / 590
Term: Winter 2012
Instructor: Jeff Howbert

Exercises 2

Date assigned: Jan. 9, 2012
Date due: Jan. 14, 2012

The math needed to answer the questions in this assignment is fairly simple. Unless explicitly requested, you aren't required to show your calculations, but it's a good idea anyway. That way, if you understand the concept, but make an arithmetic error, it's possible I can give partial credit.

1) [21 points] We have a large jar filled with balls of three colors, red, white, and black, which we will draw from the jar one at a time. There are equal numbers of each color in the jar, so the chances of drawing any particular color is 1/3. Now consider an experiment (random process) in which three balls are drawn consecutively from the jar. We record the sequence of colors as the outcome of the experiment. For example, "RRW" means the first ball was red, the second ball was red, and the third ball was white.

a) [2 points] Is the probability space for this process discrete or continuous?
b) [4 points] How many possible outcomes are there?
c) [4 points] Enumerate the outcomes.
d) [4 points] What is the probability of each outcome?
e) [7 points] What is the probability of the event "one ball of each color was drawn", where the order in which the colors were drawn does not matter?

2) [30 points] A marketing survey looked at the preferences of hot drink size among 1275 random customers of a coffee shop chain. The survey was also interested in whether the customer's gender affects their preference. The results of the survey were used to estimate the probabilities in this joint probability distribution:

	Tall (T)	Grande (G)	Veinte (V)
Female (F)	0.12	0.24	0.06
Male (M)	0.08	0.38	0.12

a) [2 points] What is p(M, T), the joint probability that a customer in the survey was both male and prefers tall drinks?
b) [4 points] What is p(F), the marginal probability that a customer in the survey was female?
c) [4 points] What is p(G), the marginal probability that a customer in the survey prefers grande drinks?

d) [6 points] What is $p(V \mid M)$, the conditional probability, given a customer in the survey was male, that he prefers veinte drinks?

e) [6 points] What is $p(F \mid V)$, the conditional probability, given a customer in the survey prefers veinte drinks, that the customer was female?

f) [8 points] There are two random variables in this situation, drink size and gender. Are they independent or dependent? Explain how you arrived at the answer, and show your calculations.

3) [8 points] What is the expected value of a single roll of a fair six-sided die? All the information you need is on Slide 12. Show your calculations.

4) [22 points] The following application of Bayes rule often occurs in actual medical practice. Suppose you have tested positive for a disease. What is the probability you actually have the disease? It depends on the sensitivity and specificity of the test, and on the prevalence (prior probability) of the disease.

We'll denote:

a positive test as	Test = pos
a negative test as	Test = neg
presence of disease as	Disease = true
absence of disease as	Disease = false

We know from clinical studies done on the test before FDA approval that the sensitivity and specificity of the test are:

p(Test = pos | Disease = true) = 0.95 (true positive rate, or sensitivity)

p(Test = neg | Disease = false) = 0.90 (true negative rate, or specificity)

From which we can also deduce:

p(Test = neg | Disease = true) = 0.05 (false negative rate)

p(Test = pos | Disease = false) = 0.10 (false positive rate)

We also know from public health surveys that the disease is relatively rare. The prevalence in the general population is:

p(Disease = true) = 0.01

From which we can deduce:

p(Disease = false) = 0.99

a) [20 points] Use Bayes rule to calculate p(Disease = true | Test = pos), i.e. the probability you actually have the disease, given the test was positive.

b) [2 points] Calculate the ratio p(Disease = true | Test = pos) / p(Disease = true). [In Bayesian statistics, a ratio like this is interpreted as the effect of new evidence on our beliefs about a probability. In this case, we are concerned with the probability of disease, and the new evidence is the test result.]

5) [21 points] Calculate the indicated products of row vector **x** and matrix **A**. You can report your results as simple row-by-row listings of elements, using tabs or spaces to separate the elements, as done below for matrix **A**.

x = [5 3 1] **A** = 0 1 2
2 1 1

a) [3 points] $\mathbf{v} \cdot \mathbf{A}^T$

b) [5 points] $\mathbf{A} \cdot \mathbf{A}^T$

c) [5 points] $\mathbf{A}^T \cdot \mathbf{A}$

d) [3 points] $\mathbf{A} \cdot \mathbf{v}^T$

e) [5 points] Explain why the answers to a) and d) are similar, using the transposition rule on Slide 52.

6) [12 points] What is the cosine of the angle between vectors **u** and **v**? Show how you set up the solution, and your actual calculations.

$\mathbf{u} = [\ 4 \quad 2 \quad 0 \quad -1\ ]$

$\mathbf{v} = [\ 3 \quad -3 \quad 2 \quad 0\ ]$