CS 188 Spring 2011	Introduction to Artificial Intelligence	Section 12 Naive Bayes and Perceptrons

Q1. Perceptrons

The Berkeley Police Department has asked you to assist with the interrogation of several suspects accused of cheating. Each suspect makes a single statement during interrogation, and from this statement you are to determine whether they are guilty (-1) or innocent (+1). To assist you, the police department provides you with records of past cheating cases: statements made by suspects, labeled with whether the suspect ended up being found guilty (-1) or innocent (+1).

You generate features from the training data as follows:

- $\phi_1(x) = n$ where "not" appears n times.
- $\phi_2(x) = m$ where "swear" appears m times.
- $\phi_3(x) = 1$, a bias term

Given a weight vector $w = (w_1, w_2, w_3)$, our classifier returns +1 if $w_1\phi_1(x) + w_2\phi_2(x) + w_3 >= 0$ and -1 otherwise.

Our training set is the following features (and the sentences they were generated from) and labels:

Training Statements	ϕ_1	ϕ_2	ϕ_3	Label
I am definitely innocent, officer	0	0	1	+1
Officer, I swear I am not lying	1	1	1	+1
I am not lying, I swear	1	1	1	+1
I am innocent, officer, I swear	0	1	1	-1
Officer, I am definitely not lying	1	0	1	-1

(a) Compute the first two updates of the Perceptron algorithm and fill in the following table, using the given initial Perceptron weights $w = (w_1, w_2, w_3)$ and data points $(\phi_1, \phi_2, \phi_3, \text{Label})$.

w	w_1	w_2	w_3
Initial	1	2	-0.5
Observing $(0, 0, 1, +1)$			
Observing $(1, 0, 1, -1)$			

(b) Will the Perceptron algorithm converge on this training dataset?

(c) Linear classifiers are often insufficient to represent a dataset using a given set of features. However, it is often possible to find new features using nonlinear functions of our existing features which do allow linear classifiers to separate the data. Nonlinear features result in more expressive linear classifiers.

For example, consider the following data set, where +'s represent positive examples and −'s represent negative examples.

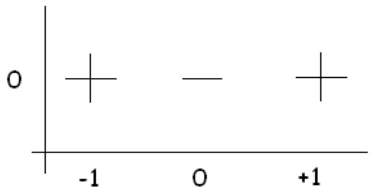

Figure 1: The original data points are not linearly separable

No linear classifier can separate the positive examples $(-1, 0)$ and $(1, 0)$ from the negative example $(0, 0)$. Rather than using a single feature, if we perform a nonlinear mapping $\phi(x_1, x_2) = (x_1^2, 1)$, the positive examples are both mapped to $(1, 1)$ and the negative example is mapped to $(0, 1)$, and we see the data can be separated by a linear classifier. One example is the line $w = [1, -0.5]$, i.e. the classifier $w^\top \phi(x) = x^2 - 0.5 >= 0$.

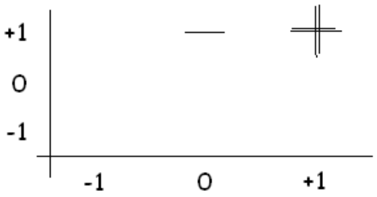

Figure 2: After the non-linear transformation the data becomes linearly separable

For what values of the weight vector $w = (w_1, w_2)$ does the classifier $w^\top \phi(x) >= 0$ achieve perfect performance on the data as shown on Figure 2? Do not use a bias term.

(d) Which of the following feature sets allows a linear classifier $w = (w_1, w_2, w_3)$ to separate the original interrogation data set? Justify your answer briefly.

(i) $\phi' = (\phi_1 + \phi_2, \phi_1 - \phi_2, 1)$

(ii) $\phi' = (\phi_1\phi_2, \phi_2^2, 1)$

(iii) $\phi' = ((\phi_1 \text{ xor } \phi_2), \phi_2, 1)$ where a xor b is 1 if either $a = 1$ or $b = 1$ but not both.

Q2. Credit Card Fraud Detection

You are building a fraud detection system for a credit card company. They have the following records of purchases for which the fraud status is known; here A is a coarse amount of a purchase, B is the kind of business purchased from, C is the country (domestic or foreign), and F is whether the transaction is fraudulent.

A	B	C	F
cheap	*candy*	*domestic*	*legit*
cheap	*jewelry*	*foreign*	*fraud*
medium	*bike*	*domestic*	*legit*
expensive	*jewelry*	*domestic*	*fraud*
medium	*bike*	*domestic*	*legit*
cheap	*game*	*domestic*	*legit*
expensive	*computer*	*foreign*	*fraud*

You decide to build a Naive Bayes classifier using these samples.

(a) Using the unsmoothed relative frequency estimates, what is the classifier's posterior distribution for the example *(cheap, computer, foreign)*?

(b) Using relative frequency estimates smoothed with add-one Laplace smoothing, what is the classifier's posterior distribution for the same example? Assume that there are no values for any of the random variables that are not present somewhere above.

(c) For add-k Laplace smoothing, what will the classifier's posterior distribution approach as $k \to \infty$? (Assume the prior distribution over classes is also smoothed.)