

CMPS 142 Fourth Homework, Winter 2010

3 Problems, 12 pts, due start of class Tuesday, Feb. 15

Each student should submit a homework and carefully acknowledge all sources of inspiration, techniques, and/or helpful ideas (web, people, books, etc.) other than the instructor, TA, and class text. The goal of this homework is to gain practice with Naive Bayes and some related techniques.

1. (4 pts) Naive Bayes.

Consider using Naive Bayes to estimate if a student will be an honor student (**H**) or normal student (**N**) in college based on their high school performance. Each instances have two measurements: the student's high school GPA (a real number) and whether or not the student took any AP courses (a boolean value). Based on the following training data, create (by hand and/or calculator) a Naive Bayes prediction rule using gaussians to estimate the conditional probability density of a high school GPAs given the class (**H** or **N**).

class	AP	GPA
H	yes	4.0
H	yes	3.7
H	no	2.5
N	no	3.8
N	yes	3.3
N	yes	3.0
N	no	3.0
N	no	2.7
N	no	2.2

Use maximum likelihood estimation (*not* Laplace estimates) for the conditional probabilities. Give the mean and variance of the gaussians you found.

Describe your prediction rule in the following form:

If AP courses are taken, predict **H** if the GPA is between ..., and
if AP courses are not taken, predict **H** if the GPA is between ...

(It is probably easier to get this description if you take logarithms.)

2. (5 pts) Bayesian decision theory.

In some classification problems we have the option to *abstain* on an instance instead of predicting a particular class. For example, if we were to classify vehicles as either **B**usses, **C**ars, or **T**rucks, we might also include an **A**bstain prediction. The loss matrix for the number of mistakes then becomes:

loss class	prediction			
	B	C	T	A
B	0	1	1	a
C	1	0	1	a
T	1	1	0	a

where a is the cost (or loss) of abstaining.

- (a) (2 pts) Assume that on some new instance we know that $P(B) = 1/2$, $P(C) = 1/4$, and $P(T) = 1/4$. What is the *risk* (expected loss) of each of the predictions **B**, **C**, **T**, and **A**? Note that the risk of **A** will be a function of a .
What values of a make abstaining have the lowest risk?
- (b) (3 pts) Formulate a general prediction rule minimizing the risk that describes when to predict **B**, **C**, **T**, and **A** as a function of $P(B)$, $P(C)$, $P(T)$ and a .
What happens when a goes to 0?
3. (3 pts) Run AdaBoost (by hand) on the following data (corresponding to the example on the slides), showing the distribution at the end of each iteration.

instance	label	h_1	h_2	h_3
x_1	-1	-1	-1	+1
x_2	+1	+1	-1	+1
x_3	+1	-1	+1	+1
x_4	+1	+1	+1	+1