CMPS 142 Second Homework, Spring 2015

4 Problems, 14 pts, due start of class Tuesday 4/28

This homework is to be done in groups of 2 or 3. Each group members should completely understand the group's solutions and *must* acknowledge all sources of inspiration, techniques, and/or helpful ideas (web, people, books, etc.) other than the instructor, TA, and class text. Each group should submit a single set of solutions containing the names and e-mail addresses of all group members. Although there are no points for "neatness", the TA may deduct points for illegible or poorly organized solutions.

1. (4 pts) Weka Experiments. The purpose of this problem is to familiarize you with Weka, gain experience with three fundamental algorithms, and to consider how changes in the data or experimental protocol are likely to affect the results. *I strongly suggest that each student do this part on their own, and combine answers for the group solutions.* Open the diabetes.arff file (in the `data` directory distributed with Weka) and read the comments at the top (lines beginning with "`%`").

   For this problem you will be running Weka's Nearest Neighbor (`IB1` in the `lazy` folder), `NaiveBayes` (in the `bayes` folder), and logistic regression (`Logistic` in the `functions` folder, see also Section 4.3.2 in Bishop) classifiers.

   (a) Select the `Use training set` test option and run the three classifiers. Report their results (accuracies). Which algorithm is best and **why**?

   (b) For your logistic regression model from the previous step, give the decision boundary as a linear formula (of the form $\sum_i w_i * x_i + bias = 0$ where i is the index of an attribute). You can find the predictions made on the data by using the "more options" button in the test options window and selecting the "output predictions" box. You can then check your linear formula by finding data points whose probability distribution is close to 0.50/0.50 and verifying that they are close to the decision boundary. (Weka outputs one logistic regression model based on the entire data set).

   (c) Repeat part a) using 10-fold Cross-validation as the test option. What changes about the accuracies? Continue to use 10-fold Cross-validation for the rest of the problem.

   (d) Use preprocessing to "normalize" the features (use the preprocess tab and select unsupervised, attribute, normalize). Read the information on this method, and look at the new attribute values. What did it do?

   Rerun the logistic regression with 10-fold cross validation and the attributes normalized. Did the accuracies change? Why?

   Are there any dramatic changes in the logistic regression weight vector? Why?

   (Note: attribute normalization is so important to nearest neighbor that IB1 normalizes the attributes automatically.)

(e) In logistic regression, the ridge parameter penalizes large weights. What happens to the cross validation accuracy and hypothesis weights when it is set to 0? How about when it is increased (to say 0.3)?

(f) Would you expect 3NN or 5NN do better than Nearest Neighbor? Why? Test your hypothesis by using `IBk` in the `lazy` folder and report the resulting accuracies.

(g) Create a modified version of the diabetes dataset by picking one attribute at random and adding 10 additional copies of that attribute to the data set (or .arff file). There should be the same number of examples, but each example will now have 19 rather than 9 attributes (including the class label). How would you expect the 10-fold cross validation accuracies of the classifiers to change? Run the classifiers on the modified .arff file and report the changes.

(h) Create a second modified version of the diabetes.arff file, this time adding 20 random (0,1)-valued attributes to the file. Re-run the algorithms on this version of the data (with pre-processing to normalize the features) and report how the accuracies changed.

2. (3 pts) Bayesian Probability.

   Probabilities are sensitive to the precise kind of information (conditioning) available.

   Assume that a new neighbor has two children, one older and one younger, and assume that their genders are like fair coin flips.

   (a) What is the outcome space and atomic events for the "experiment" that sets the genders of the two children?

   (b) Suppose I ask him whether he has any boys, and he says yes. What is the probability that one child is a girl?

   (c) Suppose instead that I happen to see one of his children run by, and that one is a boy (although we don't know if it is the elder or younger child). What is the probability that the other child is a girl?

   Hint: the answers to the two questions are different!

3. (5 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, yes=1, no=0). 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**) and a Bernoulli distribution for the AP probability . (I know that Gaussians may not fit this problem well, but use them anyway).

Recall that Naive Bayes makes the simplifying assumption that the features are conditionally independent given the class (Although in the Naive Bayes chapter Andrew Ng emphasizes fitting discrete distributions to the features, one can also fit a continuous density to the features an use the density at a feature value just like the probability under a discrete distribution), so (for example)

$$\mathbb{P}\left[\text{GPA=3.2, AP=yes} \mid \text{type=H}\right] = \mathbb{P}\left[\text{GPA=3.2} \mid \text{type=H}\right] \mathbb{P}\left[\text{AP=yes} \mid \text{type=H}\right].$$

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* the unbiased Laplace estimates) for the distributions of the two features conditioned on the two classes. Give the mean and variance of the gaussians you found for the GPA.

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, 3 digits of precision should sufficient. Also, the logarithm of the Gaussian densities are quadratic, so it is possible that two different GPA values v could both have

$$\mathbb{P}\left[\text{GPA=}v, \text{AP=yes} \mid \text{type=H}\right] = \mathbb{P}\left[\text{GPA=}v, \text{AP=yes} \mid \text{type=N}\right].$$

so the prediction rules can be finite intervals of GPA values rather than a simple threshold.)

4. (2 pt) Two random variables V and W are *independent* if for all values v and w, the events $(V = v)$ and $(W = w)$ are independent, so the joint density (or distribution) $p(V = v, W = w) = p(V = v)p(W = w)$. First prove that if V and W are independent random variables then the expectation of their product, $E[VW]$, equals the product of their expectations, $E[V]E[W]$. Assume that the random variables V and W are integer valued, so that the distributions $p(V)$ and $p(W)$ are distributions over the integers and the expectations can be written as sums.