ANLY-601 Spring 2018

Assignment 3 – Mid-term Exam
Due 5:00 pm, Monday, March 5, 2018

Please do all your own work. You may use your class notes, the primary course text, any calculus books, Mathematica (for algebraic manipulations), and your favorite numerical packages. Please do not use the internet. Please write legibly. If you use Mathematica to derive results, you must include the notebook with your solution so I can see what you did.

1. **ROC Curves — Theory** (20 points)

   Recall the likelihood ratio test for two classes

$$l(x) \;=\; \frac{p(x|\omega_1)}{p(x|\omega_2)} \;\; \begin{array}{c} \omega_1 \\ > \\ < \\ \omega_2 \end{array} \;\; \eta$$

   We defined the ROC curve as the curve traced out in $1-\mathcal{E}_1$ vs $\mathcal{E}_2$ as the threshold η is varied.

   (a) (5 points) Write the two error types $\mathcal{E}_i$ as appropriate integrals over the conditional distributions $p(h|\omega_i)$ of the the negative log likelihood ratio $h = -\log l(x)$. (The form you need is in the lecture notes.) Show that the slope of the ROC curve at the threshold value η is given by

$$\text{slope} \;=\; \frac{d(1-\mathcal{E}_1)}{d\mathcal{E}_2} \;=\; -\frac{d\mathcal{E}_1}{d\mathcal{E}_2} \;=\; -\frac{\left(\frac{d\mathcal{E}_1}{d\log\eta}\right)}{\left(\frac{d\mathcal{E}_2}{d\log\eta}\right)} \;=\; \frac{p(h=-\log(\eta)|\omega_1)}{p(h=-\log(\eta)|\omega_2)} .$$

   (b) (5 points) Show that the error for ω_1 objects can be re-written as

$$\mathcal{E}_1 = \int_{L_2} l(x)\, p(x|\omega_2)\, d^n x$$

   where $l(x)$ is the likelihood ratio (above). Note that this expression can be re-written as

$$\mathcal{E}_1 \;=\; \int_{-\log\eta}^{\infty} \exp(-h)\; p(h|\omega_2)\; dh$$

   (c) (5 points) Use the rewritten form of $\mathcal{E}_1$ from part (b) to show that the slope of the ROC is

$$\text{slope} \;\;=\; \eta \;\; .$$

   Sketch an ROC curve taking into account this result at the endpoints of the curve (i.e. at $\eta = \infty$ and at $\eta = 0$).

   (d) (5 points) Show that the ROC curve is concave downward — that is, the 2nd derivative is negative

$$\frac{d\,slope}{d\mathcal{E}_2} \leq 0 \;\; .$$

   Hint: Don't make this more difficult than it is.

2. **Parametric Classifiers and Posterior Class Probabilities** (15 points)

This exercise develops the connection between logistic regression and posterior probabilities. Consider n-dimensional input vectors $x \in R^N$ and the logistic discriminant function

$$h(x) \; = \; \frac{1}{1 + \exp -(W^T x \, + \, W_0)} \; .$$

(a) (10 points) Assume that the class-conditional densities $p(x|\omega_i)$ are multivariate Gaussian with equal covariance matrices. Use Bayes rule to write the posterior probability $p(\omega_2|x)$ in terms of the prior P_2, the class-conditional density $p(x|\omega_2)$ and the unconditional density $p(x)$. Using this expression, find the values of W and W_0 for which $h(x) = p(\omega_2|x)$.

Hint: Write the unconditional density as $p(x) = P_1 \, p(x|\omega_1) + P_2 \, p(x|\omega_2)$ where P_i are the class priors.

(b) (5 points) For this part, let $x \in R$ (i.e. a scalar rather than a vector). Sketch (or plot) the function $h(x)$ and describe how the shape of the curve $h(x)$ depends on the separation between the class means, and the variance. Why does this make sense?

3. **Bayes Classifiers for Gaussian Class-Conditional Densities**(25 points)

Construct Bayesian classifiers using the assumption that the class-conditional densities are Gaussian. The data sets are on the class Blackboard page under Assignments → Data Sets

- Pima Indians Diabetes data
- Wisconsin Breast Cancer data
- Acoustic Phoneme Data

For each, split the data in half. One half, the training set, will be used for fitting classifier parameters (class-conditional means and covariances, class priors). The second half of the data, the test set, will get used to evaluate the classifier performance. (You are not to use the test set in any phase of constructing the classifiers. It's only for evaluation.)

When you construct the training and test sets, do it in such a way that the frequency of each class is *roughly the same* in the training and test sets. For the Pima Indians and Wisconsin Brease Cancer data, I suggest that you segregate the data into two piles by classes. Then split each pile in half, one half for training, one half for test. This will insure that the class priors are about the same for the training and test set. The phoneme data is already segregated by class, you need only divide this in half for training and test. *Be sure to read the data descriptions posted with the actual data.*

(a) (5 points) The **Pima Indians** data contains 768 sample from two classes – tested positive for diabetes (268) and tested negative for diabetes (500). Each of the 768 rows in the data set contains nine values. The first eight values in each row are the features. The last value is the class label: 0 for not diabetic, and 1 for diabetic. Designate the examples labeled diabetic as ω_1 so that $1 - \mathcal{E}_1$ is the rate at which diabetic samples are correctly identified, and $\mathcal{E}_2$ is the false alarm rate.

Use the *training data* to estimate the class conditional means, covariance matrices, and class priors. Construct a likelihood ratio test to minimize the error rate and use it to classify the *test* data. Evaluate the classifier by giving the sample error rate (percent of misclassified samples) on the test set, and the empirical class-conditional error rates $\mathcal{E}_1$ and $\mathcal{E}_2$.

(b) (5 points) Repeat the exercise on the **Wisconsin breast cancer data**. Note that the first column of the data set is the case number and is not relevant for the exercise. Columns 2 through 10 are the features, column 11 has the class label (0 for benign samples, 1 for malignant samples). Designate the malignant samples as ω_1 so that $1-\mathcal{E}_1$ is the rate at which malignancies are correctly detected, and $\mathcal{E}_2$ is the false alarm rate. Construct and test the classifier as in part (a).

(c) (5 points) Finally, repeat for the acoustic phoneme data. For uniformity among your solutions, designate examples of 'n' as ω_1 and examples of 'm' as ω_2. Construct and test the classifier as in part (a).

(d) (10 points) Build ROC curves for the Pima Indians and Wisconsin Breast Cancer classifiers using the test data. To do this, vary the threshold in the likelihood ratio test away from the Bayes optimal value sweeping out the entire applicable range of the threshold for the test data. Measure $\mathcal{E}_1$ and $\mathcal{E}_2$ on the *test data* at each threshold and plot an ROC curve, $1-\mathcal{E}_1$ on the vertical axis, and $\mathcal{E}_2$ on the horizontal axis.

Also plot on each ROC curve, the single point corresponding to the classifier built in parts (a) and (b). (Note that this classifier point should be *on* the ROC curve, if not, you've done something wrong.)

4. **Linear Classifier or Logistic Regression** (25 points)

In this problem, you will build linear classifiers (or a logistic regression model, whichever you prefer) and exercise them on the same data sets as above. If you use a linear classifier, use the discriminant function

$$h(x) \;=\; V^T x \;+\; v_0 \tag{1}$$

with the vector V and the scalar v_0 given by

$$\begin{aligned} V &= (P_1\,\Sigma_1 \,+\, P_2\,\Sigma_2)^{-1}\,(M_2 - M_1) && (2)\\ v_0 &= -V^T(P_1\,M_1 \,+\, P_2\,M_2)\;\;. && (3) \end{aligned}$$

If you use logistic regression, you may build it from scratch, or use any numerical package you like.

Use the same training and test sets you used for your Bayesian classifiers. Measure and report the total misclassification error rate (as a decimal fraction, or percent), as well as the class-conditional error rates $\mathcal{E}_1$ and $\mathcal{E}_2$.

Do this for

(a) (5 points) The Pima Indians diabetes data, designating the samples "tested positive" as ω_1 so that $1-\mathcal{E}_1$ is the rate at which diabetic cases are correctly detected $\mathcal{E}_2$ is the false alarm rate. Compare your results with those in problem 3.

(b) (5 points) The Wisconsin Breast Cancer data, designating the malignant samples as ω_1 so that $1-\mathcal{E}_1$ is the rate at which malignancies are correctly detected and $\mathcal{E}_2$ is the false alarm rate.

(c) (5 points) The phoneme data.

(d) (10 points) Build ROC curves for the linear classifiers in parts (a) and (b). If you are using the linear classifier, vary the threshold v_0 away from the optimal point given by equation (3) and measure $\mathcal{E}_1$ and $\mathcal{E}_2$ on the *test data* at each threshold. Plot an ROC curve, $1 - \mathcal{E}_1$ on the vertical axis, and $\mathcal{E}_2$ on the horizontal axis.

If you use a logistic regression, you ordinarily threshold the output of the logistic function at 1/2 to assign the class of the example at the input. To use your logistic model to build an ROC curve, you sweep the output threshold throughout the full range applicable to the test data and measure $\mathcal{E}_1$ and $\mathcal{E}_2$ at each value of the threshold.

Also plot on each ROC curve, the single point corresponding to the classifier built in parts (a) and (b). (Note that this classifier point should be *on* the ROC curve, if not, you've done something wrong.)

Do this for the Pima Indians and Wisconsin Breast Cancer data sets. Compare your results with those in problem 3.