STAT 161/261: Homework 2
Bayes Decision Theory. Due Monday, April 18 in class.

Prof. Alyson Fletcher

Send all submissions and questions to the TA.

1. Maximum likelihood parameter estimation from data

   (a) Load the data `Doctor1data.txt` and `Doctor2data.txt`. The two columns are the weight (in pounds) and height (in inches) for 1000 patients at the offices of two different pediatricians in a large practice, where certain doctors have multiple subspecialties.

   (b) Plot a 2D histogram of the data from the first doctor. You can use a 2D bar plot or surface plot. In MATLAB, the bar plot can be done with the `hist3` function. Select 10 bins in each axis. Label your plot.

   (c) Assuming a Gaussian distribution, find the ML estimate for the mean and covariance matrix for the data, which are the sample mean and covariance. Plot a 2D surface plot of the bivariate Gaussian density. Qualitatively, is the Gaussian density a good fit for Doctor1's patients? Why?

   (d) Repeat parts (b) and (c) with the second data file. Is a Gaussian a good fit?

   (e) Describe a clustering and better model for the data from Doctor 2. Who do you think her patients are compared to Doctor 1's? (Note: the office specializes in infants and teenagers.)

2. The data for the problem is in the file `housePrM.txt`. Column 2 is the price of the house in thousands of dollars and column 3 is the size in hundreds of square feet.

   (a) Create a plot of price vs. size. Do you see a linear relationship?

   (b) Obtain the equation of the best linear regression fit to the data. Plot the regression fitted line on the scatterplot. What price do you predict for a house with size 20?

   (c) Do you notice anything unusual in the scatterplot?

   (d) Obtain the correlation between the size and price.

3. Use MATLAB, R or equivalent. In this problem, we will illustrate the concept of empirical risk minimization on a simple classification problem. Suppose that data (x, y) is generated from a model, where $y = 0$ or 1 is the class label with

$$P(y = 1) = P(y = 0) = \frac{1}{2},$$

and x is a scalar xith likelihood

$$p(x|y=i) = \frac{1}{\lambda_i} e^{-x/\lambda_i}, \quad x \geq 0, \tag{1}$$

where $\lambda_i = \mathbb{E}(x|y=i)$ is the conditional expectation of x given $y = i$. Take $\lambda_0 = 1$, $\lambda_1 = 10$.

(a) Derive the MAP classifier, $\hat{y}$, assuming you know the values λ_i. Generate $N_{\text{test}} = 1000$ i.i.d. samples (x_i, y_i) from the above distribution, and run the classifier on samples you generate. Measure the classification error rate, which is the fraction of samples i on which $\hat{y} \neq y$.

(b) Now suppose that you don't know the parameters λ_i. Derive the MLE of λ_i. Generate $N_{\text{train}} = 1000$ i.i.d. samples (x_i, y_i) of the above distribution, and obtain ML estimates $\hat{\lambda}_i$ from the training data. Use these estimates in the MAP classifier from part (a) and measure the classification error rate.

(c) (****Due with HW3**, but builds on earlier part of this problem.) Next suppose that the data is not exactly modeled via the exponential distributions in (1). Specifically,

$$p(x|y=0) = \frac{1}{\lambda_0} e^{-x/\lambda_0},$$
$$p(x|y=1) = \frac{1-q}{\lambda_1} e^{-x/\lambda_1} + \frac{q}{\lambda_2} e^{-x/\lambda_2},$$

so that when $y = 0$, x is distributed as before, but when $y = 1$, x is drawn from a mixture of two exponentials. Take $\lambda_0 = 1$, $\lambda_1 = 10$, $\lambda_2 = 100$, and $q = 0.1$. Generate 1000 training and test samples from this density. Then, re-do part (b), where the training and classifier still assumes that both likelihoods are exponentials. What is the classification error rate? Is this classifier robust to errors in the model?

(d) (****Due with HW3** also.) To find a classifier that is more robust, we will use empirical risk minimization. Consider a set of classifiers of the form

$$\hat{y} = \begin{cases} 1 & \text{if } x \geq \gamma, \\ 0 & \text{if } x < \gamma, \end{cases}$$

for some threshold level γ. Using the training data from the previous part, find the value of γ that minimizes the classification error rate on the training data. Then, using that value of γ, measure the classification error rate on the test data. How does this compare to the classification error in the previous part?

4. Easy-to-prove facts about linear regression.

(a) Consider the (very!) simple linear regression model $\hat{y} = \beta_0$ for data (x_i, y_i). (Estimating with a constant is equivalent to fitting a line that is constrained to zero slope.) Find the least-squares estimate of β_0 as function of N data samples.

(b) Prove that for the least squares estimator in the simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$, where ϵ is zero mean, the estimate for β_0 is unbiased.