BIOM/SYSC5405 – Pattern Classification and Experiment Design

Take-home Final — Due Wednesday 23 April.

Please work independently. Submit a single PDF file on CULearn.

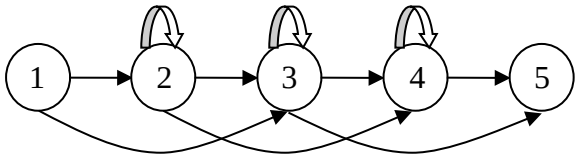
Question 1: Experiment Design

Consider the fruit data from Q1 of the assignment. We wish to create a linear discriminant classifier to differentiate between oranges and apples using only the weight feature.

a) We are considering either the Perceptron criterion or the MSE criterion to optimize the placement of our decision boundary. Which algorithm would you use to train the linear discriminant classifier, and why?
b) We want to compare various methods for estimating the true error rate of your classifier.
   i) Estimate the apparent error rate for your classifier using your algorithm from part a). Briefly describe how you would do this, and what error rate you obtained.
   ii) Use a repeated hold-out test to estimate your error rate. For each hold-out test, train your classifier on 90% of the data and test on the 10% of your data you 'held out'. Repeat this 100 times and record the mean, range, and standard deviation of your estimated error rates.
   iii) Use a repeated 10-fold cross-validation technique to estimate your error rate. Repeat a 10-fold cross-validation test 100 times (make sure you don't choose exactly the same folds each time). What is the difference between this estimator and the repeated hold-out test in part ii) above? Measure the mean, range, and standard deviation of your estimated error rates. How do these compare with your results from part i) and part ii)? Is this what you were expecting (100 words)?

Question 2: Hidden Markov Models

Consider a hidden Markov model (HMM) with the following structure:



and the following transition probabilities:

$$a_{i,j} = \begin{bmatrix} 0 & 0.4 & 0.6 & 0 & 0 \\ 0 & 0.3 & 0.4 & 0.3 & 0 \\ 0 & 0 & 0.5 & 0.2 & 0.3 \\ 0 & 0 & 0 & 0.1 & 0.9 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and the following emission probabilities (i.e. the probability that state j will emit symbol k):

$$b_{j,k} = \begin{bmatrix} 0.3 & 0.2 & 0.1 & 0.4 \\ 0.1 & 0.6 & 0.2 & 0.1 \\ 0.7 & 0.1 & 0.1 & 0.1 \\ 0.2 & 0.3 & 0.2 & 0.3 \end{bmatrix}$$

where the symbols are k_1='C', k_2='G',k_3='U',k_4='A'. Note that we always start in state 1 and that state 5 is an end state with no emissions. What is the probability of observing the sequence: 'UAG'? List all the state paths

that could produce this sequence. For a more complex model, how would you compute this probability? (describe briefly – 150 words)

Question 3: Bayesian classifier

Assume that you have measured three features for 100 samples each of two different types of fish. The data for class 1 (salmon) is found in final_Q3_data1.tsv and the data for class 2 (trout) is found in final_Q3_data2.tsv. You decide to use a Bayesian classifier. Estimate the mean vector and covariance matrix for each class. Compute the determinant and inverse covariance matrix. What is wrong? How can you fix it? (*hint, visualize your data).* Discuss what was wrong, how you fixed it, and give the apparent error rate for your classifier. (300 words, include a plot illustrating the "problem" with the feature data).

Question 4: Linear Discriminant Analysis

Assume that you have trained a generalized linear discriminant classifier resulting in the following discriminant function: $g(x) = x^t Ax + x^t b + c$ where $A = \begin{bmatrix}1 & 2\\ 3 & 4\end{bmatrix}; b = \begin{bmatrix}5\\ 6\end{bmatrix}; c = 7$.What form of classification boundary do you expect, based on the matrix A? Plot the discriminant boundary in the range x_1=[-15,15], x_2=[-15,15] (*hint*: *MATLAB's ezplot function is useful here).* Assume you had the following unlabelled data: $x_1 = \begin{bmatrix}0\\ 0\end{bmatrix}; x_2 = \begin{bmatrix}-5\\ 5\end{bmatrix}; x_3 = \begin{bmatrix}-2\\ 1\end{bmatrix}; x_4 = \begin{bmatrix}10\\ -7\end{bmatrix}$. Which point would you label next if you were using 'learning with queries'? *(discuss briefly)*

Question 5: Decision Trees

Consider the following dataset. i) What would be the root node of the decision tree using variance as the impurity measure? ii) What would be the first surrogate node if the feature tested in your root node were not available during operation? Show your work for both parts i & ii and discuss briefly. (250 words)

Colour	Shape	Weight	Taste	Class
Red	**Round**	**29g**	**Sweet**	**Apple**
Green	**Round**	**50g**	**Sour**	**Apple**
Green	**Round**	**45g**	**Sour**	**Apple**
Red	**Round**	**55g**	**Sweet**	**Apple**
Green	**Oblong**	**30g**	**Sour**	**Lime**
Green	**Round**	**41g**	**Sour**	**Lime**
Green	**Oblong**	**37g**	**Sour**	**Lime**

Question 6: Bayesian Belief Networks

You have been hired by a grocery store to accept or reject incoming shipments of bananas depending on the suspected presence of tarantula spiders. You measure the following relationships:

- 30% of your shipments come from Peru and 70% come from Mexico.
- 60% of your shipments arrive by boat and 40% of your shipments arrive by llama.
- Shipments that arrive from Peru via boat are five times as likely to have spiders as those that come from Mexico via boat.
- 40% of shipments that arrive from Mexico via llama contain spiders.
- 65% of the time, spiders form visible webs. However, shipments that do not contain spiders have been observed to contain webs 10% of the time (presumably from caterpillars).
- 40% of the time, dead monkeys are found in spider-infested shipments. 25% of the time dead monkeys are found in shipments containing no spiders (air holes...).
- All spiders are washed away when llamas swim from Peru.

- Due to the Gulf of Mexico pirates' hungry parrots, 80% of shipments that arrive from Mexico via boat contain no spiders.

Draw your Bayesian Belief Network

a) Populate your conditional probability tables
b) Compute the probability of a shipment containing spiders if it arrived by boat and had dead monkeys.
c) If you decide to reject all shipments that contain webs and dead monkeys, what percent of the time are you rejecting a ‘clean’ (i.e. spider-free) shipment?
d) The guy who inspects your bananas falls gravely ill from a spider bite and you stop inspecting any shipments. Instead, you decide to accept either all shipments from Peru, or all shipments from Mexico. Which source should you choose if you want to minimize spider bites? Which country should you choose if you want to minimize your dead monkey disposal charges?