BIOM/SYSC5405 – Pattern Classification and Experiment Design

Take-home Final — Due Tuesday 4pm 19 April.

This is an exam; please work independently. Submit a single PDF file on CULearn.

Question 1: Experiment Design

You are developing a system to predict credit card fraud. Your system should flag transactions you believe to be fraudulent such that a human operator can examine the case in detail and make a determination as to whether it should proceed. You collect transactions for 1 hour and end up with 50 examples of fraudulent transactions and 15,000 examples of legitimate transactions. You will be using an artificial neural network for your pattern classification system.

a) You need to train your network, optimize the number of hidden nodes, and test your final network. Describe how you will use the available data for these tasks. (250 words)

b) What performance metrics will you use and why? (250 words)

c) Suggest how a randomization test could be used to compare your system's performance to an existing system that randomly flags 1 in 100 transactions for human inspection? (250 words)

Question 2: Dog's Breakfast

a) What does it mean to "limit the false discovery rate" of a classifier? How would you do this? (200 words)

b) When searching for faces in an image of a crowd, some faces will be larger than others depending on distance from the camera. Discuss one feature that is invariant to scaling and one feature that is not. (250 words)

c) Prevalence-corrected precision is defined as $pcPR = \frac{Sn}{Sn+r(1-Sp)}$ where r reflects the number of negative samples for each positive sample. This is useful when you have a limited number of test samples even though the expected true class imbalance may be very high (e.g. 1000 negatives for each positive).

Assume you have a classifier which outputs scores for test samples. Simulate classification scores for 1000 positive test samples distributed as *N(50,5)* and 1000 negative test sample scores distributed as *N(60,12)*.

i) Plot the PDF of each class on a single plot.
ii) Plot the precision-recall curve
iii) On the same P-R curve, plot prevalence-corrected-precision-recall curves, assuming a true class imbalance of 10:1, 100:1 and 1000:1.

d) In the project, we assumed that protein windows that were not already known to be ubiquitinated were never ubiquitinated (i.e. negative samples); however, we expect that many of these sites are actually yet-to-be-discovered positive sites (i.e. they are ubiquitination sites, but nobody has observed this modification in the lab yet). Discuss the sensitivity of the *pcPR* performance metric (defined above) to such mislabelled negative test points, as a function of increasing class imbalance. (200 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. Rows are samples and columns are features.

You decide to use a Bayesian classifier with the assumption that the two class-conditional distributions follow multivariate normal distributions with equal priors. Estimate the mean vector and covariance matrix for each class-conditional distribution. Compute the determinant and inverse of the covariance matrices. 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 both your original and 'fixed' classifiers. (300 words, include a plot illustrating the "problem" with the feature data).

Question 4: 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:

- 45% of your shipments come from Cuba and 55% come from Mexico.
- 40% of your shipments arrive by boat and 60% of your shipments arrive strapped to trained llamas.
- 40% of shipments that arrive from Mexico via llamas contain spiders; however, when a llama swims from Cuba, all spiders are washed away 90% of the time.
- 65% of the time, spiders form visible webs. However, shipments that do not contain spiders have been observed to contain webs 5% of the time (presumably from caterpillars).
- 30% of the time, dead monkeys are found in spider-infested shipments. When a shipment is spider free, we still observe dead monkeys 15% of the time (air holes...).
- Due to the Gulf of Mexico pirates' hungry parrots, 80% of shipments that arrive from Mexico via boat contain no spiders whereas 40% of shipments that arrive from Cuba via boat contain no spiders (shorter crossing…)

a) Draw your Bayesian Belief Network

b) Populate your conditional probability tables

c) A shipment arrives by boat but you can't tell where it came from since the label was torn off by mistake. It clearly contains dead monkeys. Your longshoremen are refusing to unload the crate for fear of spider bites. Compute the probability of this shipment containing spiders.

d) If you decide to reject all shipments that contain both webs and dead monkeys, what percent of the time are you rejecting a 'clean' (i.e. spider-free) shipment?

e) 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 Cuba, 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?