Name:

Practice Midterm Exam

CS 165B – Machine Learning

May 2, 2016

9:30-10:45am

This is a **closed-book** test, with one exception: you may use one 8.5”x11” piece of paper with any notes you with to write on it (front and back).

Be sure to read each question carefully and provide all the information requested. **If the question asks you to explain, do so!**

Show your work. Write your answers in the spaces provided and, if necessary, on the back of the page. If you use the back, draw an arrow or write “SEE BACK” to make sure the graders don’t miss it. If you need more space, attach extra sheets of paper (available at the front).

Exams must be turned in by 10:45am sharp.

Good luck!

Manhattan (L1) distance: $d(x,y) = \sum_{i=1}^{d} |x_i - y_i|$

Euclidian (L2) distance: $d(x,\ y) = \|x\ - y\| = \left(\sum_{i=1}^{d} (x_i - y_i)^2\right)^{1/2}$

Minkowski (Lp) distance: $d(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{1/p}$

Laplace correction = $\frac{N_i+1}{|S|+k}$ m-estimate = $\frac{N_i+m\pi_i}{|S|+m}$

Algorithm GrowTree(D, F) – grow a feature tree from training data.

Input : data D; set of features F.
Output : feature tree T with labelled leaves.
if Homogeneous(D) **then return** Label(D);
$S \leftarrow$BestSplit(D, F) ; // e.g., BestSplit-Class (Algorithm 5.2)
split D into subsets D_i according to the literals in S;
for each i **do**
  if $D_i \neq \emptyset$ **then** $T_i \leftarrow$GrowTree(D_i, F) ;
  else T_i is a leaf labelled with Label(D);
end
return a tree whose root is labelled with S and whose children are T_i

Minority class
Imp($\dot{p}$) = min($\dot{p}$, 1–$\dot{p}$)
Gini index
Imp($\dot{p}$) = 2$\dot{p}$(1–$\dot{p}$)
Entropy
Imp($\dot{p}$) = – $\dot{p}\log_2(\dot{p})$ – (1–$\dot{p}$) $\log_2(1-\dot{p})$
√Gini index
Imp($\dot{p}$) = $\sqrt{2\dot{p}(1-\dot{p})}$

Total impurity:
$\text{Imp}(\{D_1, \ldots, D_l\}) = \sum_{i=1}^{l} \frac{|D_i|}{|D|} Imp(D_i)$

Multivariate least-squares regression (homogeneous representation)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$$

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$
$$= \boldsymbol{S}^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

PAC learning outputs, with probability at least 1–δ, a hypothesis h such that $err_D < \varepsilon$

Bayes Rule:

$$P(H_i \mid D) = \frac{P(D \mid H_i)\, P(H_i)}{P(D)}$$

False positive rate (FPR) = $\frac{FP}{N}$ = α

False negative rate (FNR) = $\frac{FN}{P}$ = β

True positive rate (TPR) = $\frac{TP}{P}$

True negative rate (TNR) = $\frac{TN}{N}$

Accuracy = $\frac{TP+TN}{P+N} = \left(\frac{P}{P+N}\right) TPR + \left(\frac{N}{P+N}\right) TNR$

Error rate = $\frac{FP+FN}{P+N}$

Precision = $\frac{TP}{\hat{P}}$

Ranking classifier error rate: *rank-err* = $^{err}/_{PN}$

Ranking classifier accuracy: *rank-acc* = 1 – *rank-err*

Sample Midterm Questions

NOTE: My answers are rather brief – you may need to explain your answers more thoroughly than is done here!

1. [4 points] What is machine learning? What are the three key components of a learning problem?

Machine learning is the design and analysis of algorithms that improve their performance at some task with experience.

2. [2 points] Give an example of an unsupervised discrete learning problem.

Clustering

3. [4 points] What are the three data sets that are typically used in developing a machine learning algorithm or application, and how is each used?

Training (to train ML models), validation (to choose among the alternative models), and testing (to evaluate the chosen model's performance),

4. [4 points] Describe inductive learning and deductive learning. Which do we focus on in machine learning and why?

See 3/28 lecture notes, slide 9

5. [4 points] Briefly discuss the relationship between overfitting and generalization in machine learning. What is likely to have lower error on the training data, a linear model or a higher-order polynomial model? How about on the testing data?

See 3/30 lecture notes, slides 7+; overfitting leads to poor generalization. A linear model will generally have higher error on the training data and lower error on the testing data (because it doesn't overfit the training data).

6. [3 points] What is the intrinsic dimensionality of a set of data?

The real (and relevant) dimensionality of the problem without noise or irrelevant data.

7. [6 points] A test for a new, deadly strain of anthrax (that has no symptoms) is known to be 99.9% accurate. The chances of any random person having this strain are one in

a million. You get tested for anthrax during a routine medical exam, and your test comes back positive. If A is the variable that describes whether you have anthrax (true) or not (false), and T is the variable that describes the output of your anthrax test (true if you test positive, false if you test negative), what is the relatively likelihood that you have anthrax? Use Bayes' Rule.

Given:
P(T=true | A=true) = P(T=false | A=false) = 0.999
P(A=true) = 0.000001
Find:
Relative likelihood of P(A=true | T=true) and P(A=false | T=true)

P(A=true | T=true) = kP(T=true | A=true)P(A=true) = k(0.999)(0.000001) = k(.000000999)
…where k = 1/P(T=true)
P(A=false | T=true) = kP(T=true | A=false)P(A=false) = k(0.001)(0.999999) = k(0.00999999)
So it's about 10,000 times more likely that you don't have anthrax.
(Although, note that before you got tested it was 1,000,000 more times likely that you don't have anthrax. So your chances have gone up.)

8. [3 points] Give an example of a predictive machine learning task and an example of a descriptive machine learning task.

Predictive ML – classification, regression
Descriptive ML – clustering

9. [4 points] In the basic binary linear classifier (with the linear discriminant function midway between the class centroids), the centroid of our positive class is at (2, 1, 4) and the centroid of our negative class is at (4, 4, 6). The decision boundary is a plane defined by the vector **w** and the threshold t. If we now add a new positive sample at (0, -2, 2) to the training data and recompute, how will this affect the placement of the decision plane?

It will move toward the positive class, since the new sample is on the far (positive) side of the centroid of the positive class. (Thus it will move the centroid further away from the negative class, bringing the halfway decision boundary with it.)

10. [3 points] You are given the probability tables for P(data | hypotheses) and P(hypotheses). You need to choose the best model (hypothesis) from this data. What kind of decision rule is this?

This is a maximum a posteriori (MAP) decision rule, since you can calculate kP(hypotheses | data) from what is given and choose the hypothesis that maximizes this.

11. [4 points] Why is feature selection often performed in a machine learning problem before learning the model?

To reduce dimensionality, eliminate unneeded features, decorrelate features…

12. [3 points] In testing the binary classification model you learned from the training data, you get 70 out of 100 instances correct. 40 of those correctly estimated the positive class (the concept), and the rest correctly estimated the negative class. There were a total of 60 positive examples in the test set. What is the false positive rate? The false negative rate? The precision?

TP = 40, TN = 30, FP = 10, FN = 20, FPR = 10/40 = 0.25, FNR = 20/60 = 0.33, precision = 40/50 = 0.80

13. [4 points] In learning a classifier, you use a loss function in weighing the effects of various training data instances. If the classification margin of an instance is very high, what should the loss function for that instance be (qualitatively, not a specific value)? If the margin of an instance is very low (negative), what should the loss function for that instance be? If an instance is badly misclassified, what should the loss function for that instance be?

For a large margin, it should be very low or zero. For a very low (very negative) margin, it should be high (maybe not too high to avoid bad effect of outliers). A badly misclassified instance is the same situation as the previous answer, for a very low margin.

14. You need to estimate prior probabilities for your 5-class classifier, and you have { 10, 8, 14, 9, and 5 } samples from your classes. What are your estimated prior probabilities, using Laplace correction?

11/51 = 0.22, 9/51 = 0.18, 15/51 = 0.29, 10/51 = 0.20, 6/51 = 0.12

15. Describe how a loss function (for classification or regression) can be made robust to outliers.

Instead of increasing monotonically as the margin becomes more negative, increase at first (to penalize errors) but then decrease (to not penalize large errors too much, assuming they are errors that should have minimal impact).

16. In concept learning, what is the difference between a *hypothesis* and a *concept*?

Nothing, they're the same thing. Or in some terminology, a hypothesis is an estimate of the (true) concept.

17. Give an example of a hypothesis that is not learnable by the conjunctive hypothesis space representation.

H : feature1=A or feature2=B (CHS cannot learn disjunctions of features)

18. In our conjunctive hypothesis space learning, we generally seek the least general generalization. Compared with a more general generalization, what effect does choosing a less general hypothesis have on our false positive rate?

It reduces the false positive rate – less general hypothesis means it's less likely to label something as belonging to the concept. Thus there will be fewer non-concept instances mistakenly classified as belonging to the concept.

19. What CHS hypothesis will guarantee a *consistent* concept for any problem?

The concept $h(x)$ = False.
This is guaranteed not to allow any false positives, and thus is consistent. (Recall the definitions of complete and consistent.)

20. In a decision tree learning routine, a particular node has 8 examples of the positive class and 2 examples of the negative class. What is the impurity measure of that node?

0.72, using the entropy measure

21. A decision tree for concept *c* has five leaves with the following training examples in each leaf:
    L1: (5 pos, 2 neg)
    L2: (6 pos, 1 neg)
    L3: (3 pos, 4 neg)
    L4: (0 pos, 3 neg)
    L5: (1 pos, 2 neg)

    Using Laplace correction, give the ranking order of the leaves (from highest ranked to lowest).

Empirical probabilities:
L1: 6/9 = 0.67
L2: 7/9 = 0.78
L3: 4/9 = 0.44
L4: 1/5 = 0.2
L5: 2/5 = 0.4

Rank (high to low): L2 > L1 > L3 > L5 > L4

22. In a univariate linear regression problem, what is the geometric interpretation of the regression coefficient?

The slope of the regression line.