Exam 1 Review Topics

Data preprocessing
- Data exploration
- Data cleaning
- Feature engineering
- Scaling
- Dimensionality Reduction

Decision Trees
- Gini & Entropy
- Selecting the best split based on gain
- Using gain ratio instead of gain
- Decision boundaries & characteristics

Cross Validation and Overfitting
- Cross validation process: how to do it and why to do it
- Overfitting: what it is
- Pre-pruning (Chi-squared pruning will not be on exam)
- Post pruning with a validation set: REP algorithm
- Model selection with a validation set, including nested cross-validation

KNN
- Computing distance & Voting
- Choosing k / model selection / nested cross-validation
- Decision boundaries & characteristics

Naive Bayes
- Using Bayes Theorem to make a prediction
- Laplace smoothing
- Decision boundaries & characteristics
- (there will NOT be a text classification problem on the exam)

Evaluating Classifiers
- Error/Accuracy & issues with using them
- Confusion matrices: how to make them and how to understand/use them
- TPR, FPR, TNR, FNR
- Precision, Recall, F-measure
- Issues with a class imbalance & ways to mitigate a class imbalance
- Using a cost matrix to evaluate a classifier
- Using a cost matrix to make a prediction

Ensembles
- Why ensembling is done & understand general concept of ensembling
- Know how each method works: Bagging, Boosting, Random Forest, Multi-class partitioning, Stacking

SVMs

- Conceptually understand what an SVM is doing - how it finds the optimal hyperplane
- Difference between hard-margin SVM and soft-margin SVM
- Understand the parameters of a soft-margin SVM (C and slack variables)
- How to use SVMs with multi-class problems, with categorical variables, and with non-linearly separable data
- Decision boundaries & characteristics

Neural Nets

- Forward propagation, including using a given activation function
- Calculating error
- Conceptually understand gradient descent
- Types of gradient descent (batch, stochastic, mini-batch)
- (The calculus of backpropagation will NOT be on the exam)

Practice Exam

The following practice exam includes questions on major topics covered in the class. It also gives you an idea of the format of the exam.

This practice exam is <u>not</u> indicative of the length of the exam. The in-class exam is designed to be completed in a 1hr 15 min class period.

This practice exam is not necessarily comprehensive of every topic. You will still want to review the slides and all of your notes from class.

You will be given the following formulas for the exam. If something is not on this formula list, you need to memorize it.

$$Entropy = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Where c is the number of classes;
p_i is the fraction of records belonging to class i;
and 0 log2 0 = 0 in entropy calculations

$$\log_2 X = \frac{log_{10} X}{log_{10} 2}$$

$$Gain_{split} = Impurity(parent) - \sum_{i=1}^{k} \frac{n_i}{n} Impurity(i)$$

$$GainRatio_{split} = \frac{Gain_{split}}{SplitInfo}$$

$$SplitInfo = -\sum_{i=1}^{k} \frac{n_i}{n} log_2 \frac{n_i}{n}$$

Where k is the number of splits and n_i is the number of records in partition i

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

$$\min \frac{||w||^2}{2} + C\left(\sum_{i=1}^{N} \xi_i\right)$$

Subject to $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) >= 1 - \xi_i$

Practice Exam 1: Classification

Part 1

After the first exam in a data mining course, the results of the exam were recorded along with some information about each student. The data is below:

ID	Passed All Assignments	GPA	Language	Passed Exam
1	No	3.1	Python	Yes
2	No	2.0	Python	No
3	Yes	3.5	C++	Yes
4	Yes	2.5	Java	Yes
5	Yes	3.9	Python	No
6	No	2.9	C++	No
7	Yes	3.2	Java	Yes

1. Using a KNN classifier with K=3, predict whether the following student will pass the exam. (Do not worry about normalizing the data.)

8	Yes	3.0	C++	?

2. Using a Naive Bayes classifier, predict whether the student will pass the exam. Bin the GPA feature into >=3.0 and <3.0

3. If you want to create a decision tree to classify the data, what is the best attribute to split on first?

- **Bin the GPA feature into >=3.0 and <3.0.**
- **Use Gini index as the measure of impurity**
- **(Also know how to use entropy as the measure of impurity, either one is fair game for the exam!)**
- **Because 'Language' can be split 3 ways, use Gain Ratio to compare split quality**

Part 2

Given the following cost matrix and the confusion matrices for two different classifiers:

Cost Matrix		Predicted	
		+	-
Actual	+	-1	20
	-	2	0

Classifier 1		Predicted	
		+	-
Actual	+	50	20
	-	130	300

Classifier 2		Predicted	
		+	-
Actual	+	60	10
	-	30	400

1. Which classifier is better on the basis of error rate?

2. Which classifier is better on the basis of F-measure (for the positive class only)?

3. Which classifier is better on the basis of cost?

4. Given a KNN classifier with k=7, and a new record to classify, the KNN classifier finds that of the 7 nearest neighbors to this new record, 6 of them are in the negative class, and 1 of them is in the positive class. Using the cost matrix above, what would you classify this new record as, based on Risk?

Part 3 - Short Answers

1. What is the difference between noise and outliers?

2. Give 2 ways of dealing with missing values in a dataset.

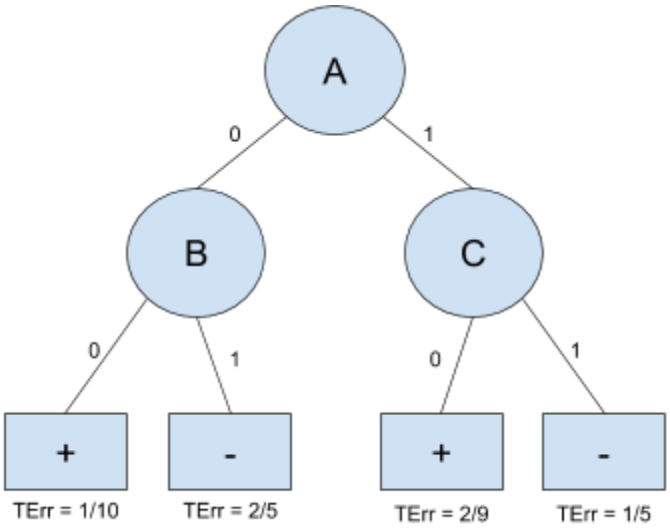
3. What is the curse of dimensionality?

4. Explain the difference between bagging and boosting.

5. Describe and/or draw a situation in which using unweighted voting for KNN gives you a different classification than weighted voting.

Part 4

1. Given the following decision tree and validation data, which nodes would be pruned with the Reduced Error Pruning algorithm? The error rate from the training data is noted at each leaf node.

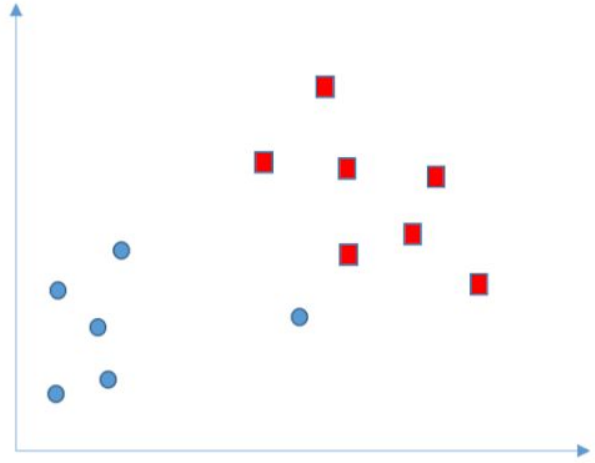


Validation set:

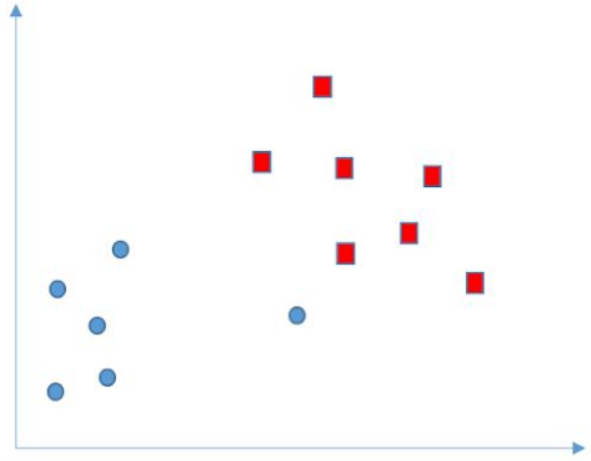
A	B	C	Class
0	0	0	+
0	1	1	+
1	1	0	-
1	0	0	-
0	1	0	+

Part 5

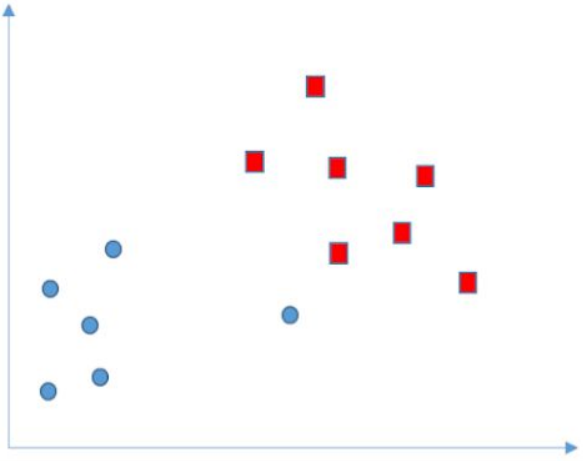
1. Draw a hyperplane (including the margin lines) generated by a linear hard-margin SVM. Circle the support vectors.



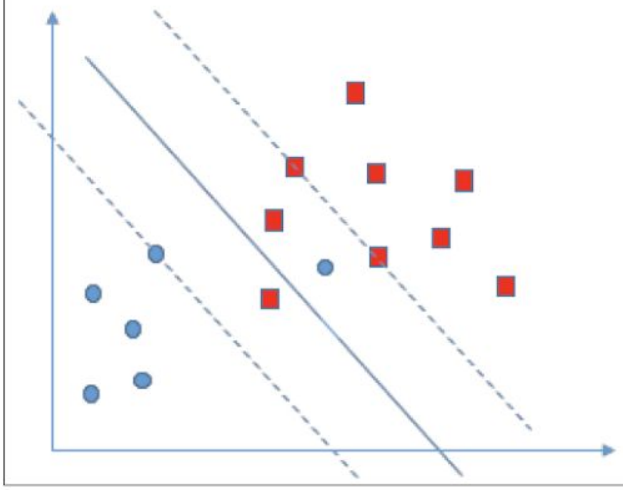
2. Draw a hyperplane (including the margin lines) generated by a linear soft-margin SVM with a large value of C (i.e. $C \rightarrow \infty$). Circle the support vectors.



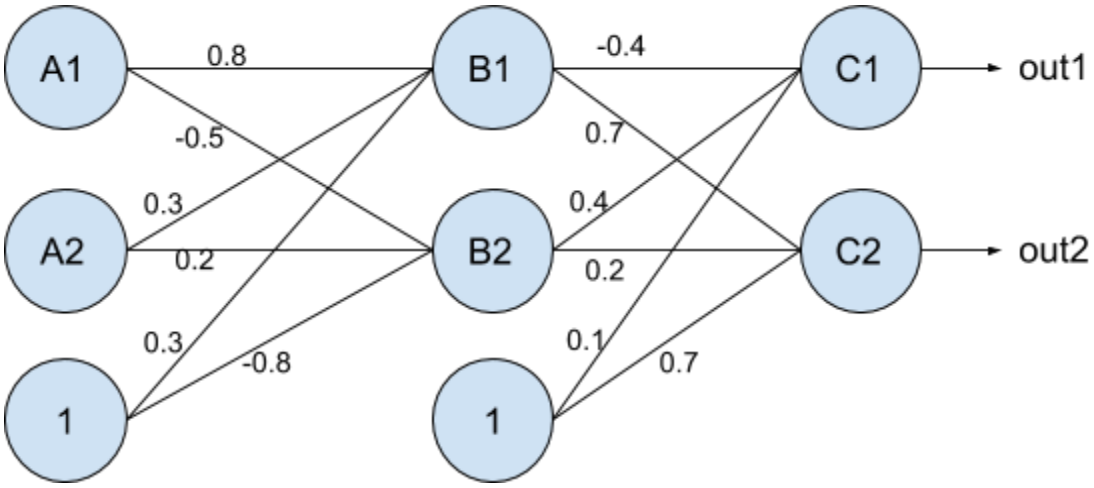
3. Draw a hyperplane (including the margin lines) generated by a linear soft-margin SVM with a small value of C (i.e. $C \rightarrow 0$). Circle the support vectors.



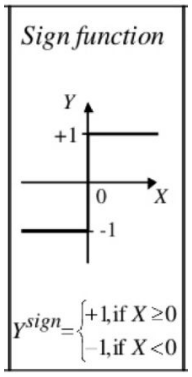
3. Draw a circle-class data point with slack $0 < \xi < 1$



Part 6



1. For the above neural network, out1 represents Class A and out2 represents Class B.
All nodes use the sign function as their activation function.



Show the forward pass of the following record. How would this neural network classify this record (Class A or Class B)?

Feature1	Feature2	Class
5	10	?