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
- Gain ratio and split info
- Decision boundaries & characteristics

Linear Regression
- Simple linear regression:
- Understanding the least-squares method to find best fit
- Calculating parameters for simple linear regression
- Regression evaluation metrics - how to calculate each & understanding meaning of each: SSE, MSE, RMSE, MAE, R2
- Multiple linear regression: understand concept (you will NOT need to calculate parameters of multiple linear regression)
- Handling non-numeric features

Cross Validation and Overfitting
- Cross validation process: how to do it and why to do it
- Overfitting: what it is
- Model selection/hyperparameter tuning with a validation set, including nested cross-validation: how to do it and why to do it

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

Evaluating Classifiers
- Error/Accuracy & issues with using them
- Confusion matrices: how to make them and how to understand/use them
- Precision, recall, F-measure, TP, FP, TN, FN
- Issues with a class imbalance & ways to mitigate a class imbalance

Naive Bayes
- Using Bayes Theorem to make a prediction
- Laplace smoothing
- Decision boundaries & characteristics

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 objective function of the SVM (both hard and soft margin)
- 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
- The kernel trick
- Decision boundaries & characteristics

Non-linear regression

- Know how each works: regression trees, KNN regressor, support vector regressor (at a basic/high level understanding)

Practice Exam Problems

The following are practice exam problems on major topics covered in the class.

This is a sample. These practice problems are not comprehensive of every topic. You will still want to review the slides, all of your notes from class, and the practice problems. These are just examples to give you an idea of the format of the exam.

This practice exam is not necessarily indicative of the length of the exam. The exam will be designed to be completed in a 1hr 15 min exam period.

=====

If you do NOT bring a cheat sheet, this formula sheet will be provided to you:

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

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

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

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

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

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

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Exam 1 Practice Problems

Part 1 - Classification Algorithms

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. Given the following dataset:
Different tissue papers & whether or not they are good for your science experiment.
(Yes, the color matters in this problem.)

ID #	Color	Acid Durability	Strength	Class
1	Yellow	7	7	bad
2	White	7	4	good
3	Yellow	3	4	good
4	Green	1	4	good
5	White	5	5	bad
6	White	6	3	bad

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

- **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!**

Note: This problem is too long for an exam, so I won't ask you to do something this long on the exam. But you do need to know how to do this - it'll just be something shorter on the exam.

SVMs: Make sure you understand the SVM practice problem questions!

Part 2 - Linear Regression

A scientist is researching whether or not birds exposed to pollutants lay eggs with thinner shells. She collects a sample of egg shells from 5 different nests and measures the pollution level and thinness of the shell. Her results are below:

Pollution	3	8	30	25	15
Thinness	1	3	9	10	5

1. Find the equation of the regression line for this data.

2. Calculate the R2 of the line.

3. Calculate the RMSE of the line.

Part 3 - Evaluating Classifiers

Given the following confusion matrices for two different classifiers:

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)?

Part 4 - 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. What is overfitting and why is it a problem?

5. What is the naive assumption in Naive Bayes?

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

7. Explain “slack” in an SVM - what is it and why do we need slack variables?