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 1	3.1 0.01	Python 2	Yes
2	No 1	2.0 1	Python 2	No
3	Yes 0	3.5 0.25	C++ 0	Yes
4	Yes 0	2.5 0.25	Java 2	Yes
5	Yes 0	3.9 0.81	Python 2	No
6	No 1	2.9 0.01	C++ 0	No
7	Yes 0	3.2 0.04	Java 2	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++	?

Yes

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

$P(Yes|Yes) \cdot P(>=3.0|Yes) \cdot P(C++|3.0) \cdot P(Yes)$

$3/4 \quad 3/4 \quad 1/4 \quad 4/7 = \frac{9}{16} \cdot \frac{1}{7}$ = Yes

$1/3 \quad 1/3 \quad 1/3 \quad 3/7 = \frac{1}{9} \cdot \frac{1}{7}$

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.

$imp(p) = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = 0.5$

W G Y

$W = 1 - \left(\frac{1}{3}^2 + \frac{2}{3}^2\right) = \frac{4}{9}$

$G = 1 - (1^2) = 0$

$Y = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = \frac{1}{2}$

$Split \rightarrow \frac{1}{2}\left(\frac{4}{9}\right) + 1(0) + \left(\frac{1}{3}\right)\left(\frac{1}{2}\right) = \frac{2}{9} + \frac{1}{6} = \frac{7}{18}$

$Gain = 1 - 7/18 = 1/9$

Acid

1	3	5	6	7	7
G	G	B	B	G	B

$<4 = 1 - (1^2) = 0$

$>4 = 1 - \left(\frac{3}{4}^2 + \frac{1}{4}^2\right) = 1 - \left(\frac{10}{16}\right) = 3/8$

$4 = \frac{1}{3}(0) + \frac{2}{3}\left(\frac{3}{8}\right) = \frac{1}{4}$

$Gain\ 4 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$

$>6.5 = 1 - \left(\frac{1}{2}^2 + \frac{1}{2}^2\right) = \frac{1}{2}$

$<6.5 = 1/2$

$6.5 = \frac{2}{3}\frac{1}{2} + \frac{1}{3}\frac{1}{2} = \frac{1}{2}$

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

$Gain\ 6.5 = 1/2 - 1/2 = 0$

Stregth

3	4	4	4	5	7
B	G	G	G	B	B

$<3.5 = 1 - (1^2) = 0$

$>3.5 = 1 - \left(\frac{3}{5}^2 + \frac{2}{5}^2\right) = 1 - \left(\frac{13}{25}\right) = \frac{12}{25}$

$3.5 = 0 + 5/6\left(\frac{12}{25}\right) = 2/5$

$Gain\ 3.5 = \frac{1}{2} - \frac{2}{5} = \frac{1}{10}$

$<4.5 = 1 - \left(\frac{3}{4}^2 + \frac{1}{4}^2\right) = \frac{3}{8}$

$>4.5 = 1 - (1^2) = 0$

$4.5 = \frac{2}{3}(3/8) + 0 = \frac{1}{4}$

$Gain\ 4.5 = 1/2 - 1/4 = 1/4$

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.

$\bar{Y} = 28/5 \quad \bar{X} = 81/5$

$\beta_1 = 0.330 \quad \beta_0 = 0.259$

2. Calculate the R2 of the line.

$$\frac{var(m) - var(F)}{var(m)} = \frac{11.84 - 0.724}{11.84} = 0.9388$$

3. Calculate the RMSE of the line.

0.8509

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?

$\frac{150}{500} = 30\%$ $\qquad$ $\frac{40}{500} = 8\%$

✓ ↗

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

$p = \frac{50}{180}, \frac{60}{90}$ $\qquad$ $r = \frac{50}{70}, \frac{60}{70}$

$$F_1 = \frac{2 \cdot \frac{50}{180} \cdot \frac{50}{70}}{\frac{50}{180} + \frac{50}{70}} \qquad F_2 = \frac{2 \cdot \frac{60}{90} \cdot \frac{60}{70}}{\frac{60}{90} + \frac{60}{70}}$$

$\downarrow$ $\qquad$ $\downarrow$

$2/5$ $\qquad$ $3/4$

✓ →

Part 4 - Short Answers

1. What is the difference between noise and outliers?

random errors and variations

Unusual data

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

Imputation, Deletion

3. What is the curse of dimensionality?

more dim → more sparse

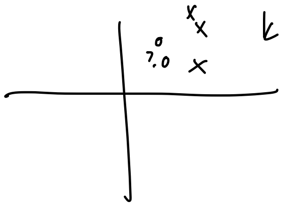
4. What is overfitting and why is it a problem?

Low training error but high test error

5. What is the naive assumption in Naive Bayes?

independence

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



k=5 → X unweighted

→ o if weighted

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

To allow misclassifications,

Reduce overfitting