NAME: ID:

CS 412/512 Machine Learning
Midterm 1

100pt

Nov. 2, 2017

- *Allocated space should be enough for your answer. Give **brief & clear explanations** for full credits. Points will be taken off for irrelevant/rambling information given within an anser.*
- *Please write legibly and **circle your final answer.***
- *You **may make additional assumptions** if you think it is necessary, but if you do so, **clearly state them**. Your grade will depend on whether the assumption was necessary.*
- ***Show your work for full credit!***
- *No Internet, no cell phones, no calculators!*

Question		***Score***	***Max Score***
1	***Basic Concepts***		***20***
2	***Decision Trees***		***24***
3	***Probability***		***16***
5	***Pdfs***		***12***
6	***Bayesian Classifiers***		***25***
7	***MLE estimate***		***3***
TOTAL			***100***

NAME: ID:

1) 20pt – Basic Concepts

a) 4pts – In a regression problem, you have trained a system to approximate a mapping from x to y. What is the **mean square error** of the estimate (f'(x)), over the given test set? *Show your work.*

Labelled Test (x,y)	Estimate f'(x)
x= 2 , y = 10	8
x= 5 , y = 15	16
x= 5 , y = 14	16
x= 6 , y = 16	19

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\frac{1}{4}\left((2)^2 + (1)^2 + (2)^2 + (3)^2\right)$$

$$4 + 1 + 4 + 9 = 18/4 = 9/2$$

MSE = 4.5

b) 10pts - You have a training set, a test set and a learning algorithm (for instance a decision tree). Answer as true/False (Note: statements without a qualifier (generally, often etc) claims to hold in general; so choose T if it does indeed.). 2pt each answer. -1 each wrong guess.

- **T / F** A zero training set error indicates good generalization performance.
- **T / F** A system that has higher test set error compared to its training set error has overfit to the training set.
- **T / F** As the number of features increases, the risk of overfitting generally increases.
- **T / F** As the number of training samples increases, the risk of overfitting generally decreases.
- **T / F** More complex models with larger number of parameters may fit the training data well, but they are more likely to overfit compared to smaller models.

c) 6pts – Consider a classification problem with two possible output labels (classes C1 and C2) such that one class (C1) has a 0.9 prior probability.

- **3pts -** What is the **base error rate** for this problem, indicate as a percentage? Hint: ZeroR from Weka.

  Majority $\frac{9}{10}$ so $\frac{1}{10}$ = %10 base error rate

- **3pts –** What would be the expected error rate if you pick a label randomly (you select C1 and C2 each with a probability of 0.5) for a given x; indicate as a percentage?

  $$\frac{5}{10}\cdot\frac{9}{10} + \frac{1}{10}\cdot\frac{5}{10} = \frac{45+5}{100} = \%50$$

NAME: ID:

2) 24pt – Entropy, Decision Tree Learning

Given:

x	$\log_2 x$
0.25	-2
0.33	-1.6
0.5	-1
0.66	-0.6
0.75	-0.4
1	0

a) 3pt – What is the entropy of a random variable **dice** that represents the output of a **4-sided** (possible outputs are 1,2,3,4) **fair dice? Show your work.**

1 2 3 4

$-\frac{1}{4}\log_2 2^{-2} - \frac{1}{4}\log_2 2^{-2} - \frac{1}{4}\log_2 2^{-2} - \frac{1}{4}\log_2 2^{-2}$ $\quad \frac{1}{2}\cdot 4 = 2$

-2

$1/2$

b) 3pt – How would the entropy in a) **change** if the dice was **biased** (for example, the probability of having a 1 is higher than 2,3, 4)? It would;

1/4 1/4 1/4 3/4

- decrease
- increase
- remain unchanged

$-3/4 \log_2^{3/4}$

$-0.4 \cdot \frac{-3}{4} = +0.3$ so decrease

Circle the appropriate answer. No explanation necessary. -1 pt for wrong answer.

c) 4 pts – Fill-in-the-blanks or answer as true/False, as appropriate. 2pt each. -1pts off each false guess.

top-down

- **T/F** The **greedy** decision tree learning algorithm ID3 that we saw in class is **optimal** in the sense that it always generates the **smallest** tree (least number of nodes). not always

- State **one of the most important advantages** of using a decision tree classifier.

Can be used for both classification and regression problems ✓

Fast and efficient ✓

Normalization is not required ✓

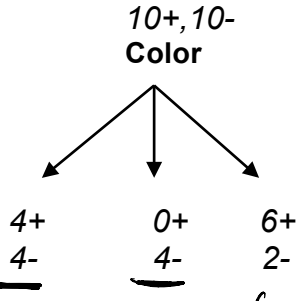
Non parametric less data preparation needed ✓

NAME: ID:

d) 10pt – What is the **remaining entropy of class labels** after the tree is split according to the feature "**Color**". The +s represent one class, and –s represent another class. There are a total of 20 samples (10+,10-) at the root of the tree.

NOTE: **if the answer is very simple (e.g. entropy is 0 or 1), you can just write that down without writing the whole formula.**

Entropy: randomness or measuring the disorder of the information



2pt - Entropy at left leaf: ..... $-\frac{1}{2}\log_2^{-1} - \frac{1}{2}\log_2^{-1}$ = 1

2pt - Entropy at middle leaf: ..0.....

3pt - Entropy at right leaf: ..... $-\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$ (−0.4) $= \frac{-3}{14}\cdot\frac{-4}{10} - \frac{1}{14}\cdot -2 = \frac{3}{10} + \frac{1}{2} = \frac{2}{10}$ (5) 0.2

3pt - Remaining Entropy = ...

$-\frac{1}{2}\log_2^{2^{-1}} - \frac{1}{2}\log_2^{1^{-1}} = 1$?

e) 4pts – Your boss gave you a regression problem and asked you to use decision trees with the ID3 algorithm, but the training set size is small.

- **2pts -** How woud you evaluate different trees, using a **validation set or cross-validation**? Give a one-line explanation for your reason.

→ In ID3 alg. iteratively divide features. Use entropy and information gain as metrics.
→ Two params characterize d.t. those we learn by fitting the tree and those we set befor training
cv => how to set the hyper params to increase performance of the resulting tree as much as possible

- **2pts -** After you are done with training different models and measuring error on validation data, what would you then give to your boss as the **finished system**? Give a one-line answer.

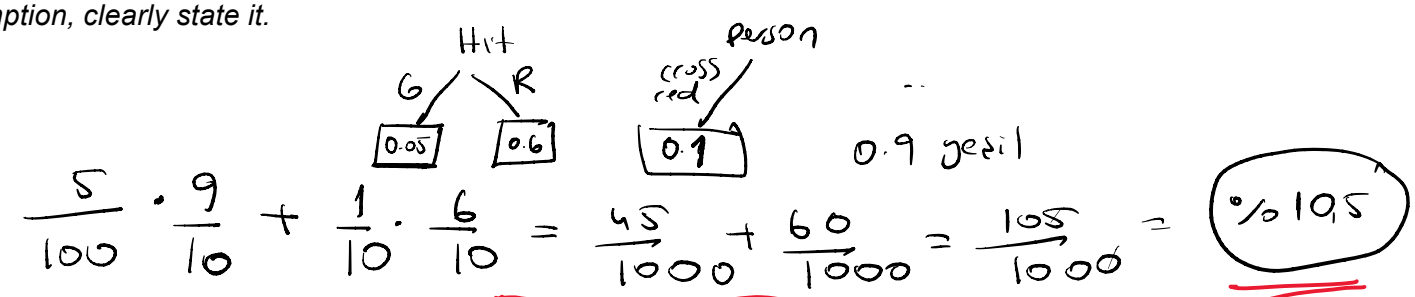
→ Present decision tree models with the comparision of error rates with diff. models.

NAME: ID:

3) 16pt – Probability Theory

a) 4pt – A pedestrian can be hit by a car with a low probability (p=0.05) when crossing the road when the light is green light for pedestrians. The probability of being hit by a car while the light is red to pedestrians is expectedly high (p=0.6). There is no yellow light in this problem.

What is the total probability of being hit? You should assume that most persons will be reasonable and only get tempted to cross the road at red light with a low probability (p=0.1). *If you must make an assumption, clearly state it.*



b) 4pts – Answer the following based on the **joint probability table** involving two random variables X and Y, given below. Show your work (do not just give a single number).

P(A|B) = P(B|A)P(A) / P(B)

	Y=Red	Y=Green	
X=1	0.1	0.0	0.1
X=2	0.1	0.4	0.5
X=3	0.3	0.1	0.4
	0.5	0.5	

i) P(X=2)= ..... 0.5 / 1 = 1/2 ..... 2pt

ii) P(Y=Red | X=2) = ... P(X=2 | Y=Red) · P(Y=Red) / P(X=2) 2pt

or 0.1 / 0.5 = 2/10

0.2

(0.1/0.5) × 1/2 / 0.5

(1/10) / (1/2) = 1/5 = 0.2

c) 4pts –Assume that two random variables A and B are **independent**. **Simplify the following probabilities** (probability terms should be simpler probabilities, involving fewer terms).

- P(A, B) = ... P(A) · P(B)

dependent ise P(A) · P(B|A)

- P(A | B) = ... P(B|A) P(A) / P(B)

P(B)

P(A)

For independent events P(A|B) = P(A)

d) 4pts –Assume that two random variables X, Y are **conditionally independent given C**. Simplify the following probabilities using the conditional independence information. Hint: Above question ☺

- P(X, Y | C) = ... P(X,Y) = P(X) · P(Y)

- P(X | Y, C) = ... P(X)

P(Y,C)

P(Y,C|X) · P(X) / P(Y,C)

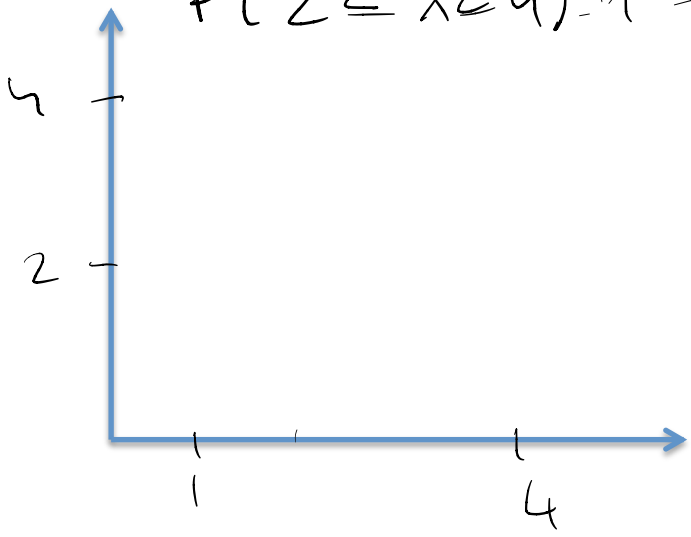
P(X)

NAME: ID:

5) 12pt – PDFs

Assume p(x,y) is distributed uniformly in the rectangular area between x in [1-4] and y in [2-4] and 0 elsewhere.

a) 4pts – Draw the pdf p(x,y) making sure to label axes (similar to what we did in our homework)



$P(2 \le X \le 4) = 1 = \frac{1}{2} \times 2$

Find area $= 3 \cdot 2 = 6$

$$\int_2^4 \int_1^4 p(x,y)\,dx\,dy = \int_2^4 \int_1^4 \frac{1}{6}\,dx\,dy = 1$$

uniform distribution
$p(x,y) > 1$ conform the rule
integral $= 1$
$P_2 = \int_A p(x,y)\,dx\,dy \in [0,1]$
for some area subset
$A \subseteq [1,4] \times [2,4]$
$p_2(x,y) = 1$ defined on $[1,2] \times [2,3]$

b) 4pts – What is the value of the density (p(x,y)) for (x,y) inside the rectangular region?

c) 4pts – What is the marginal probability of P(2 <= x <= 3)?

NAME: ID:

6) 25pt – Bayesian Decision Theory

Consider a **classification problem** with input **x** and k possible classes C_i **for the questions a)-d).**

a) 3pt - State the **Bayes formula** that relates **prior, posterior and conditional probabilities** of a class C_i, given some input **x**. *One line formula.*

$$P(C_i|x) = \frac{P(x|C_i)\,P(C_i)}{P(x)}$$

posterior = likelihood × prior

b) 3pt – What is the **Bayesian decision criterion that minimizes misclassification error (i.e. to which class do you assign a given x)?** One line formula, but do not skip details in the formula.

To be able to minimize misclassification error, we need to choose the one with has higher posterior probability

$$1 - E(\max_i Pr(y=i|x))$$

$$E_{bayes} = 1 - \sum_{k=1}^{k} \int_{C_k} P(C_k)\,P(x|C_k)\,dx$$

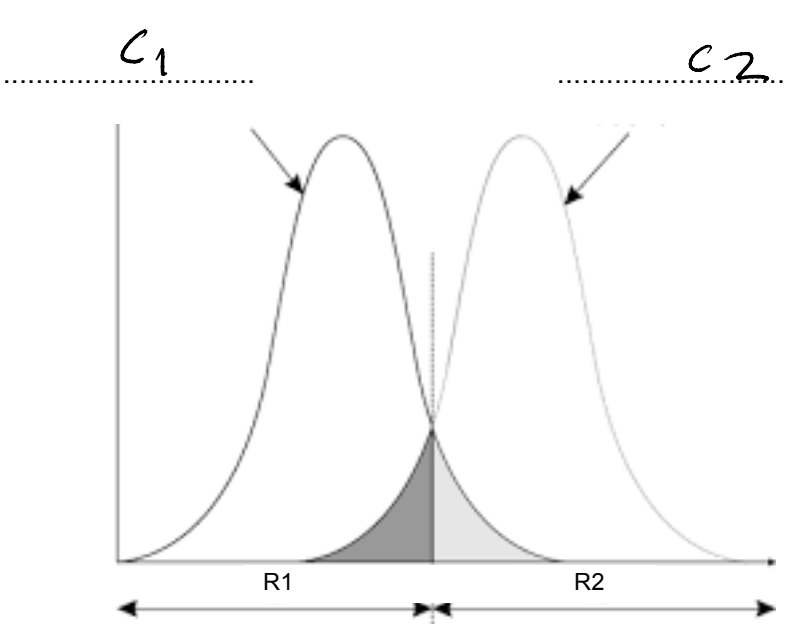
c) 3pt – Assume we are given x= [a1 a2 ... ak] where a_i are the attributes C_j is the jth class. How is the term below simplified, if we assume the Naive Bayes classifier. Be careful about details/indices....

P(a1,a2,..,ad | Cj) = ... $C_{NB} = \arg\max_{C_j \in C} P(C_j)$

$$\arg\max_{a_i \in a} P(C_j)$$

NAME: ID:

d) 6pt – Assume we have the following two distributions for x for the two classes C1 and C2.



- **3pts - –** Write the appropriate labels on each of the two distributions, so that the given decision boundary is optimal (minimizes misclassification error). **Be careful, no partial.**

- **3pts –**Indicate the area that corresponds to the probability of error corresponding to C2 samples being classified as class C1 and give its formula as an integral.

NAME: ID:

f) 10pt –

- **6pts** - Consider a naive Bayes classifier trained on the dataset given below. A new patient comes **who has x=[High Fever and Body Ache, but NO runny nose, and NO throat pain].** Calculate only the posterior probability of having Flu given these symptoms without considering the denominator (P(symptoms)). Do not use smoothing for this example.

You should just leave as a product of probabilities, without doing the final arithmetic.

	Fever	Body Ache	Runny Nose	Throat Pain	Disease
1	High	Yes	No	Yes	Flu
2	High	Yes	No	No	Flu
3	High	No	Yes	No	Flu
4	Medium	Yes	No	No	Flu
5	Medium	No	No	No	Flu
6	High	Yes	No	Yes	Flu
7	Low	No	Yes	Yes	Common cold
8	Low	No	Yes	Yes	Common cold
9	Low	Yes	No	No	Common cold
10	Medium	No	Yes	Yes	Common cold

$P(Flu \mid High, Yes, No, No) = \frac{P(High, Yes, No, No \mid Flu) \cdot P(Flu)}{P(High, Yes, No, No)}$

use Fever bc it has higher probability

- **4pts** – Use Laplace smoothing to calculate only:

P(Fever=High | Flu) = ...

P(RunnyNose=No | Flu) = ...

$P(Flu \mid High) \cdot P(high)$

$P(Flu)$

NAME: ID:

7) 3pts –

Assume you have observed N coin tosses and 8 of them are Heads and 2 are tails.

- What is the ML estimate of the probability p of observing a head with this coin?

N coin toss

Head

$\frac{8}{N}$ $\frac{2}{N}$

8 Head

2 tail

$\frac{8}{N} = p(\text{head})$