NAME: ID:

CS 412/512 Machine Learning
Final

100pts

Jan 2, 2018

- **Allocated space should be enough for your answer. Give brief & clear explanations for full credits.**
- **Please write legibly and circle your final answer.**
- **No questions. You may make additional assumptions if you think it is necessary, but if you do so, clearly state them. You will get full points if the assumption was indeed needed.**
- **No Internet, no cell phones, no calculators.**
- **Whenever applicable, you are expected to use proper mathematical notation.**

Question		*Score*	*Max Score*
1	*Basic Concepts*		*25*
2	*Multivariate Gaussian-Bayes*		*20*
3	*Probability*		*10*
4	*Neural Networks*		*10*
5	*Classifier Combination*		*12*
6	*General*		*22*
TOTAL			*100*

NAME: ID:

1) 25pts – Assume we have the following **gender detection problem**, given the observed hair length and height (**both discrete** random variables) that are observed from a distance. Answer the following questions according to the data given below. Use proper probability notations. Note sample index is just there to facilitate counting, it is not an attribute.

Sample Index	**Hair Length**	**Height**	**Gender**
1	Short	160	Male
2	Short	170	Male
3	Short	170	Male
4	Short	180	Male
5	Medium	160	Male
6	Medium	170	Male
7	Long	170	Male
8	Medium	160	Female
9	Long	160	Female
10	Long	170	Female

a) 2pt – What is the **prior probability** for Female?

b) 2pts - What is the error rate for **random guessing**? Give as a percentage.

c) 2pts – What is the **base error rate** for this problem? Give as a percentage.

d) 2pts – What is the probability that the person is Male if you observe that the person has Height = 170?

e) 2pts – What is the **joint probability** of seeing a Male with long hair?

f) 2pts – What is the **intrinsic error** for this problem? (1- maximum accuracy a classifier can have for this problem)?

NAME: ID:

Sample Index	Hair Length	Height	Gender
1	Short	160	Male
2	Short	160	Male
3	Short	170	Male
4	Short	180	Male
5	Medium	170	Male
6	Medium	170	Male
7	Long	170	Male
8	Medium	160	Female
9	Long	160	Female
10	Long	170	Female

Same table repeated here for your convenience.

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

Table for entropy calculations if necessary

g) 5pts – Draw the decision tree for this problem, making sure to use the **most informative** (one that reduces the remaining entropy the most) attribute at the root. No need to show your entropy calculations, but do show the label or probability associated with each leaf.

h) 4pts – How would a **Naive Bayes** classifier **that does not** uses smoothing classify a given person with Short hair and Height= 160? State the decision and show your work, but no need to do the final numeric calculation.

i) 4pts – Explain why **Laplace smoothing** is desirable in general? One line answer. Also show how the probability changes for Hair length and the Female class for this example.

2) 20pt – Multivariate Normal – Bayes Classifier

Assume we have a **K-class** classification problem where the data consists of **10 real valued attributes** that are assumed to be distributed according to a 10D **multi-variate Normal distribution** with parameters **μ_k and Σ_k,** separately for each class C_k. Answer the following according to this setup.

a) 1pts – What is the size (dimensions) of the covariance matrix Σ_k?

b) 4pts – How many **independent** (i.e. not counting those that can be derived from the others) **parameters** are there to estimate for one class? (Consider both μ_k and Σ_k**).** Give the exact number for 10 attributes. **How about for K=10 classes?**

For one class: ....

For K=10 classes: ....

c) 2pts – How many independent parameters are there to estimate **in total** if we assume classes share a common covariance matrix Σ?

For K=10 classes: ....

e) 3pts – The shared covariance matrix assumption is often wrong, yet it still works better than not making the assumption in most cases. Explain in only one line each as to when it should be used and why it often performs better.

When to use?: ....

Why it is often useful? ....

NAME: ID:

f) 4pts – True/False (2pt each) – -1 for each wrong answer.

- **T / F** If P(X,Y) is 2D Normal, then X and Y are 1D normal (i.e. the individual dimensions of a joint normal distribution is also normally distributed).

- **T / F** If X and Y are each 1D normal, then P(X,Y) ~ 2D Normal (i.e. if two random variables are normally distributed, so will be their joint distribution).

Reminder for multivariate Normal distribution:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

g) 6pts – Give the formula for the discriminant function $g_k(\mathbf{x})$ for class C_k and multi-dimensional input **x**, assuming a **common covariance matrix Σ shared between the classes**.

Simplify as much as possible and be careful about details in the formula.

$g_k(\mathbf{x}) = \log P(C_k \mid \mathbf{x})$

= ....

= ....

NAME: ID:

3) 10pt – Probability

a) 6pts - In the general population, one in 1,000 people have the dreaded X disease. There is a test for this disease, but unfortunately, it is only 99% accurate. That is, if you have the disease, 99 times out of 100, the test will turn out positive; if you do *not* have the disease, 99 times out of 100 the test will turn out negative.

> **You take this test and the results indicate you have the disease. Use Bayesian reasoning to compare the probabilities that you actually have the disease or not.**

Note: Leave your answer as a product of the component probabilities without actually doing the arithmetic.

P (.................... | ....) =

vs.

P(............................|......................) =

b) 4pts - A continuous random variable $x \in \mathbb{R}$ has a uniform probability density between $-a \leq x \leq a$, and zero elsewhere. Draw the pdf and state what is the density p(x) in the interval [-a,a].

p(x) = ...

4) 10pt – Neural Networks

a) 2pts – Draw the graph of the logistic (sigmoid) function f(n)= = $1/(1+e^{-n})$.
Be careful and complete with your labels.

b) 4pts – The derivative of the sigmoid f’(n) can be factored as a product of two terms (shown on the right):

$$\boldsymbol{f'}(n) = \frac{d}{dn}\left(\frac{1}{1+e^{-n}}\right) = \frac{e^{-n}}{(1+e^{-n})^2} = \left(1-\frac{1}{1+e^{-n}}\right)\left(\frac{1}{1+e^{-n}}\right)$$

- What **simplicity** does this formulation bring in neural network training with sigmoidal activation functions? One line answer. Hint: Think of the forward propagation.

- What does **’sigmoid saturation’** refer to in the context of neural networks? One line answer.

c) 4pt – Give the *architecture, weights and biases* of a single node *network* that takes a **binary input X** and returns **NOT(X).** In other words:

If X is 0, the output should be 1.
If X is 1, the output should be 0.

Your bias **should work** with the hard-limiter (threshold) activation function $f(x) = 1$ if $x \geq 0$
$= 0$ otherwise

NAME: ID:

5) 12pts - Classifier Combination

a) 8pts – Fill-in-the-blank – 2pts each.

- **2pts – What is the most important advantage in combining classifiers?** One-line answers only.

- **2pts** – Bagging has the disadvantage of **not** ....training data.

- **2pt –** Boosting trains the next classifiers to ....

b) 6pts –Define the concepts of **bias** and **variance** by considering an estimator $\hat{\theta}$ of θ, calculated over the given data set X. Use expectation over different data sets X. One line answers only!

(Mathematical notation) **Variance**: ....

(Verbal explanation) **Variance**: ....

(Mathematical notation) **Bias:** ....

(Verbal explanation) **Bias:** ....

6) 22pts - General

i) 8pts - True/False (1pt each, but -1pt for each wrong answer.

Note: If there is no qualifier (generally, often, ...) it means the statement is claimed to always hold.

- **T / F** Any regression problem can be learned with zero training error, using a decision tree with arbitrary number of nodes.

- **T / F** If you are given M data points, and use half for training and half for testing, the difference between training error and test error decreases as M increases.

- **T / F** The Bayes decision boundary found using equal class priors would move towards the more likely class if class priors were not equal.

- **T / F** As the number of hidden nodes increases, the risk of overfitting generally increases in neural networks.

ii) 10pts - Fill-in-the-blanks.

- **2pts - Entropy** of a random variable ranges between ………………………and……………………..

- **2pts - Weight decay** is a form of .... used in neural networks.

- **2pts -** Decision trees are preferred in applications such as law because the rules they use in the

decisions are …………………………

- **4pts -** Scaling the attributes is very important for some classifiers. Given a training set that consists of 2D samples x = $[x1\ x2]^T$, give the formula for transforming the data to be **zero-mean, unit variance**. Be careful about details.

$\mathbf{x} = \begin{bmatrix} x1 \\ x2 \end{bmatrix} \longrightarrow z = \begin{bmatrix} z1 \\ z2 \end{bmatrix}$ where z1 = ...

z2= ....

NAME: ID:

YOU NEED TO ANSWER ONLY 4PTS-WORTH OF QUESTİON FROM THİS PAGE.

CHOOSE TO ANSWER ONLY ONE OF THE FOLLOWING (iii or iv) BY WRITING SKIPPED IN THE OTHER ONE.
(If not selected, one question will be chosen randomly)

iii) 4pts – Explain what regularization is and why is it useful in 2 lines. You can use neural networks or polynomial curve fitting as context for your explanation if you wish.

iv) 4pts – You are going to classify mammographies as cancerous (positive) or not, using **inbalanced** (few positives) training data and we are mainly interested in reducing the false negatives (estimating that a cancer case is not).

How would you approach the problem? State 2 things that may help and be clear. You may comment about data preparation, preprocessing, loss functions or what is a suitable classifier.

- ....

- ....