NAME: ID:

CS 412/512 Machine Learning
Midterm 2

99pt

Dec. 12, 2017

- *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 please. You may make additional assumptions** if you think it is necessary, but if you do so, **clearly state them**.*
- ***No Internet, no cell phones!***

Question		*Score*	*Max Score*
1	*Multivariate Normal*		*15*
2	*Neural Networks*		*30*
3	*Classifier Combination*		*15*
4	*SVM*		*15*
5	*General*		*25*
TOTAL			*100*

Reminders:

Sigmoid(x) = $1/(1+e^{-x})$

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

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

NAME: ID:

1) 15pt – Multivariate Normal – Bayes Classifier

a) 4pts – Answer as True/False, as appropriate. 2pt each. -1 each false guess for T/F questions.

- **T/F** The correlation between two random variables x, y is symmetric (i.e.. $\rho_{xy} = \rho_{yx}$).

- **T/F** The maximum likelihood estimate μ_{ML} of the mean μ of a Normal distribution is unbiased.

b) **6** - Fill-in-the-blank:

- 3pts - What is the correlation between the two random variables for the given $covariance\ matrix$

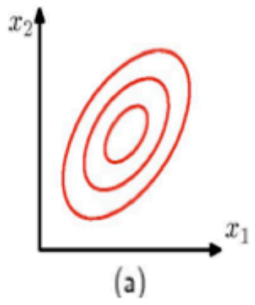
  $\sum = [\ 4 \quad -6\ \ -6 \quad 9\ ]$. **Answer**:

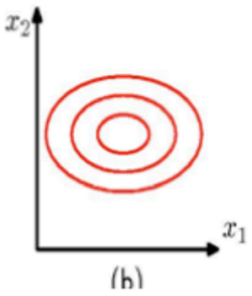
  .....

- 3pts - "The maximum likelihood estimate σ_{ML} of the standard deviation σ of a Normal distribution is biased because

  ..... ≠.....". *(Be careful to details)*

c) 4pt – Give suitable covariance matrices for the two Normal distributions that are shown with their equal density contours. No need to worry about which variation may be bigger.





NAME: ID:

Σ = **Σ=**

2) 30pt – Neural Networks

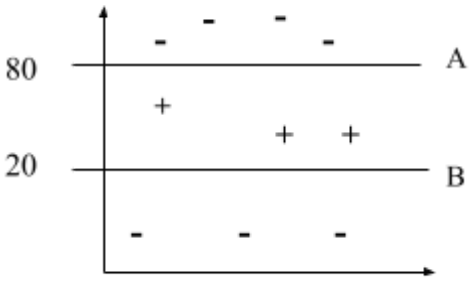
a) 6pt –You are looking for the **minima** of the following function $f(x,y) = y^2 – 5x + 2xy$. Starting with the point x= 1 and y= 1, trace **one step of the backpropagation algorithm** by computing the derivative and **finding the next value for the (x,y), using a step size of 0.1**. *Show your work.*

Gradient =

Next values for (x,y) = ...

b) 8pt – We are given the following binary classification problem, where **+ is the positive class and – is the negative class.**

Draw them on the figure as well.



$\mathbf{w_A}$ = $\mathbf{w_B}$ =

i) 4pt –What is the **weight vectors w_A and w_B,** corresponding to the following two decision boundaries A and B, respectively?

ii) 4pts – Give the *architecture, weights and biases* of the *full network* that uses the above weight vectors and can classify a given **input (x,y)** appropriately, as belonging to the + or – class.

NAME: ID:

c) 4pts – Consider a neuron that takes **N binary inputs** and uses the threshold activation function. **What is a suitable value for its bias b, in order for the neuron to be active if k or more of its N inputs are 1?** Assume all weights are 1.

b = ...

d) 4pt - Assume we are dealing with a <u>regression</u> problem. Complete the following derivative that shows how the squared error $E_p=(t_p - o)^2$ for a pattern p ,with target t_p, changes with changes in weight w_i of the **output unit**. o is the output of the system. Make sure to show your derivation. Hint: Oneline differentiation.

$\boldsymbol{\delta E_p/\delta w_i =}$

NAME: ID:

e) 8pts - True/False (2pt each) – -1pt for each wrong answer.

- **T / F** A neuron with a saturated sigmoid activation (close to 1 or 0) will learn very quickly.
- **T / F** Linearly non-separable problems (e.g. XOR) can be solved with two layers of *linear* units (neurons with linear activations).
- **T / F** A shortcoming of gradient descent based methods, such as backpropagation, is that they may get stuck in local minima.
- **T / F** Neural networks are susceptible to (affected by) scale differences among the different dimensions (attributes) of the input.

3) 15pts - Classifier Combination

a) 4pts – Fill-in-the-blank or Answer as True/False as appropriate (2pt each).

- 2pts - Consider en ensemble which outputs the arithmetic average of the outputs of some k base classifiers. Compared to the base classifiers, the ensemble is **expected** to have lower ..... ? (circle all correct choice)

  a) bias b) variance c) both

- 2pts – Bagging works by generating different ..... for each of the base classifiers.

NAME: ID:

b) 4pts – Considering a **majority vote** ensemble on a two class problem, assume that each of the 5 base classifiers has a probability of error *p* on a given input and that their errors are independent. What can you say about the probability of error for the ensemble? *Check all that apply.*

- P(ensemble makes error) <= P(exactly 3 base classifiers make errors)
- P(ensemble makes error) >= P(exactly 3 base classifiers make errors)
- P(ensemble makes error) = P(exactly 3 base classifiers make errors)
- Cannot say

d) 7pts – Consider Error Correcting Output Codes with K classes (C1..Ck) and L base classifiers (h1..hL). Assume you have 100 samples from each of the 3 classes in your training set. Answer the following according to the code matrix given below.

	h1	**h2**	**h3**	**h4**
C1	+1	+1	-1	-1
C2	-1	+1	-1	+1
C3	+1	-1	+1	+1

i) 2pts - What is the task assigned to first base classifier (h1)?

ii) 2pts - What is the training set and its size, for the first base classifier (h1)?

NAME: ID:

iii) 3pts – Assume the 4 dichotomizers give the output **[-1 -1 +1 +1]** for a given input x. How would you classify x? Show your work.

NAME: ID:

4) 15pt – SVM

a) 6pt –What are support vectors and what is the margin in Support Vector Machines? Explain them in words and to clarify add a simple 2-dimensional problem.

b) 4pts – Consider a soft margin SVM with the parameter C for penalizing instances inside the margin (or on the other side). State the effect of C on the size of the margin (how large or small):

As C increases, ... because

.....

c) 5pts – Consider the kernel $K(x,y) = 5(\mathbf{x}.\mathbf{y}+1)^2$ for $\mathbf{x},\mathbf{y}$ in R^2

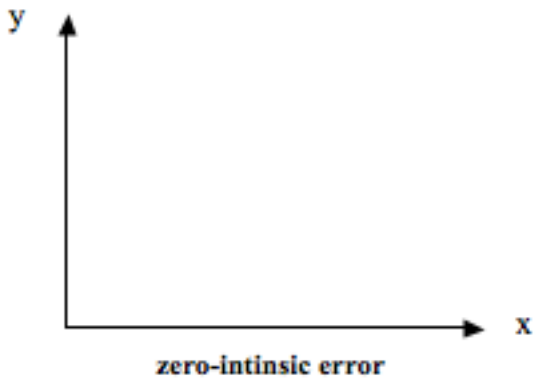
Note: (**.** is the dot product; **x** = [x1 x2]; kernel is 5 times (x dot y plus 1 squared)).

Show **$\varphi(x)$ in the correspondence of $K(x,y) = \varphi(x).\ \varphi(y)$,** where $\varphi(x)$ is the **implicit mapping** of x into the higher-dimensional space $z=\varphi(x)$. I.e. what are the dimensions of z in terms of the dimensions of x. Hint: *First expand the kernel.*

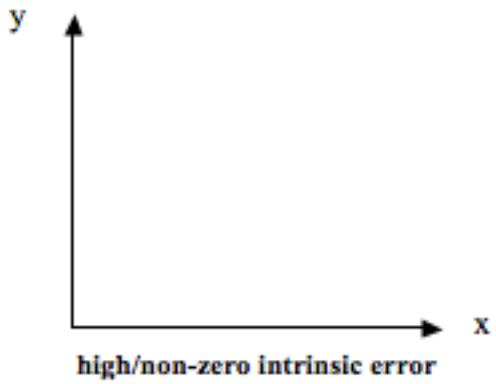
NAME: ID:

5) 25pts – General Concepts

a) 4pts - Show your understanding of the concept of **intrinsic error,** by drawing a sample data set $X=\{x^1,...x^N\}$ along with the target values y^i, **such that in the first case there is zero-intrinsic error** and **in the second case there is a high intrinsic error.** *You must be clear to get full point, so draw enough details to explain the concept.*



zero-intinsic error



high/non-zero intrinsic error

b) 4pts – Label the terms in the following formula where **y(x; D) is the estimated mapping for x, *learned from training set D*; ED indicates the expectation aken over different data sets; and h(x) is the actual mapping that the learner is trying to estimate.**

$$\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D})-h(\mathbf{x})\}^2\right] = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]-h(\mathbf{x})\}^2}+\underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D})-\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}.$$

Answer: …….. ……..

NAME: ID:

c) 4pt – Given the following data where each data point is given as a tuple (x,y=f(x)), what is the **conditional expectation of y given x=4**; i.e. E[y|x=4]?

(x1,y1) (x2 y2) . . . (x6 y6)
{ (2,50), (4,30), (4,30), (4,40), (6,4), (10,150) }

E[y|x=4] = …….

d) 4pt –One can take two different approaches to classification: **discriminative** versus **generative.** Write the type of each of the following classifiers next to their name.

i. Decision Trees ……..
ii. Bayesian Classifier ……
iii. Neural Networks ……
iv. SVM ……

e) 4pt – Define **likelihood** considering a training data set $X=\{x^i\}$ and some distribution parameter θ?

$L(\Theta)$ = ... = ...

By definition **Assuming N i.i.d data $x^i \in X$**

NAME: ID:

f) 2pts – What is the Mahalanobis distance from **a point p** to a Gaussian distribution with mean μ and covariance matrix Σ? Give the formula and be careful to the details!

g) 3pts –Assume a random variable X is sampled from a one dimensional normal distribution with mean μ and standard deviation σ, answer the following accordingly. What is the probability that X takes on a value larger than μ?

- P(x >= μ) = ...