CSCI-B555 Midterm Exam -- 100 Points

Spring 2015 **Prof. Predrag Radivojac**

Question

Name: Enrique Areyan

Closed book, closed notes, open mind.

Remember:

1) Each question is worth 10 points.

2) You are to draw a line through the 2 (two) questions you do NOT want counted.

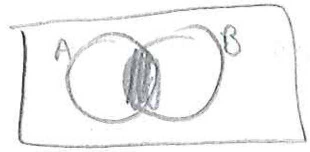
3) Be sure to show all your work.

4) Do all the work on these exam pages.

5) Do not take exam apart.

6) If you need more space, use the back of the adjoining page and tell me where to look!

7) **Good Luck!**



$\frac{4}{6} + \frac{1}{6}$

$\frac{5}{6} - 1$

Problem 1. Miscellaneous

1.1. (4 points) Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A, B \in \mathcal{F}$. Using only the axioms of probability and basic set operations, prove or disprove that $P(A \cap B) \geq P(A) + P(B) - 1$.

$P(A) + P(B) - 1 = P(A) + P(B) - P(\Omega)$ ✓, since $P(\Omega) = 1$

$= P(A) + P(B) - P(A^c \cup A)$ ✓

$= P(A) + P(B) - P(A^c) - P(A)$ ✓

$= P(B) - P(A^c)$ ✓

$\geq P(A \cap B)$ ... By Axioms. ? I don't see it.

2/4

1.2. (2 points) What is the main difference between supervised and unsupervised learning?

In supervised learning you have positive and negative examples (or labels) in your data which you use to create a learning algorithm, whereas in unsupervised learning you don't have labels. ✓

2

1.3 (2 points) Write the perceptron update rule. Assume that $\mathbf{x} \in \mathbb{R}^k$ and $y \in \{-1, +1\}$.

if example x_i is missclassified

if $w^T x_i \leq 0$ but $y_i = +1$ ← underclassifed

$w^{(T+1)} = w^{(t)} + x_i$ ✓

else if $w^T x_i > 0$ but $y_i = -1$ ← overclassified

$w^{(T+1)} = w^{(t)} - x_i$

2

1.4 (2 points) Briefly describe the difference between batch and stochastic (incremental) modes of optimization.

1) In batch mode of optimization we loop through the whole data set (i.e., we have a finite set of points) AND perform operation. We can also go back and work through the same data AGAIN.

Not quite...

1/2

2) Stochastic mode is more like online optimization where we have a data stream where each data point is received once, AND then discarded. In a sense there is no "going back".

Problem 2. Elements of Probability Theory

2.1 (4 points) Consider a measurable space $(\Omega, \mathcal{F})$, where $\Omega = [0,1]$ and $\mathcal{F} = \mathcal{B}(\Omega)$. Define a set function P on this space as follows

$$P(A) = \begin{cases} 1/2 & \text{if } 0 \in \mathcal{F} \text{ or } 1 \in \mathcal{F} \text{ but not both} \\ 1 & \text{if } 0 \in \mathcal{F} \text{ and } 1 \in \mathcal{F} \\ 0 & \text{otherwise} \end{cases}$$

Is P a probability measure? Show your work.

P is a probability measure if $P(\Omega) = 1$ and if $A \cap B = \Phi \Rightarrow$ $P(A \cup B) = P(A) + P(B)$.

$[0, \frac{1}{2}$

2.2 (4 points) Consider a probability space $(\Omega, \mathcal{F}, P)$ and any two events A and B from $\mathcal{F}$. Using axioms of probability and elementary set operations ($\cup$, $\cap$, and complement), prove that:

a) (2 points) $P(A \cap B^c) = P(A) - P(A \cap B)$

b) (2 points) $P(A^c \cap B^c) = 1 - P(A) - P(B) + P(A \cap B)$

2.3. (2 points) Let of (X, Y) be a discrete random vector. Write a definition of the conditional expectation $E[X|y]$ if $p_{XY}(x, y)$, the joint probability mass function, is known.

$$E[X|Y] = \sum_{x \in \Omega_x} x \cdot p(x|y)$$

$$= \sum_{x \in \Omega_x} x \cdot \frac{p(x,y)}{p(y)} = \sum_{x \in \Omega_x} x \cdot \frac{p(x,y)}{\sum_{x \in \Omega_x} p(y,x)} = \frac{1}{\sum_{x \in \Omega_x} p(x,y)} \sum_{x \in \Omega_x} x \cdot p(x,y)$$

Problem 3. Random Variables

3.1. (5 points) Independence and conditional independence.

a) (2 points) Provide a mathematical formulation of conditional independence between two random variables X and Y, given some other random variable Z.

$$\Pr\{X=x, Y=y \mid Z=z\} = \Pr\{X=x \mid Z=z\} \cdot \Pr\{Y=y \mid Z=z\}$$

$\forall x \in \Omega_x, y \in \Omega_y, z \in \Omega_z$

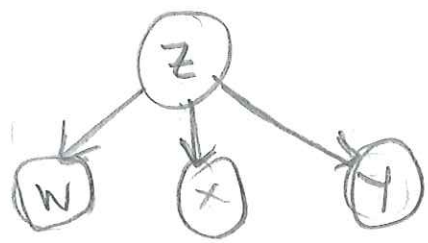
2

b) (1 point) True or false. Independent random variables are conditionally independent regardless of the variable they are conditioned on.

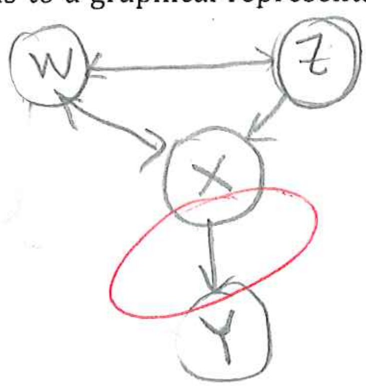
1 FALSE

c) (2 points) Suppose you are given four random variables W, X, Y, and Z. Variables W, X, and Y, are conditionally independent given Z. Provide a graphical representation of this situation.

2



3.2 (3 points) Suppose we are given four random variables W, X, Y, and Z and their joint probability distribution $P(W,X,Y,Z)$. If we know that $P(Y|X) = P(Y)$, provide a factorization of $P(W,X,Y,Z)$ that corresponds to a graphical representation with the smallest number of edges. Use directed graphs.



0/3

3.3 (2 points) Suppose X is a random variable such that $E[X] = 2$ and $E[X^2] = 8$. Calculate $E[(2+4X)^2]$.

$$E[(2+4x)^2] = E[4 + 16x + 16x^2]$$
$$= E[4] + E[16x] + E[16x^2] \quad \Big\} \text{ linearity of expectation}$$
$$= 4 + 16E[x] + 16E[x^2]$$
$$= 4 + 16 \cdot 2 + 16 \cdot 8 = 4 + 32 + 128 = \boxed{164}$$

2

Posterior $\propto$ like
$P(M|X)$ $P(X)$

Problem 4. Foundations of Classification and Regression

4.1 (2 points) Define Bayes risk classifier or intuitively explain what it is.

$$\int_M c(M,\hat{M})\, p(\hat{M}|D)\, dM$$

(this becomes a sum when model is discrete)

the Bayes risk is the risk we incur when selecting model $\hat{M}$ when M is the true model.
We want to minimize this risk, i.e., get as close as possible to the true model.

½/2

4.2 (2 points) What is the optimal regression model when the cost (or loss) function between the prediction $f(\mathbf{x})$ and the true target value y is $c(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$ for any data point?

The mean of the posterior distribution

$E[y|x]$ ✓

2

Squared error loss $\Rightarrow$ optimal regression = mean of posterior distribution
(minimization of risk as in 4.1)

4.3 (2 points) Briefly explain the difference between generative and discriminative classifiers.

generative: we learn a function from the data as in we generate a model.

discriminative: we discern data points from positive and negative examples to come up with the model.

½/2

4.4 (2 points) What is the main difference between multi-class and multi-label classification?

multi-label means that a data point might belong to one or more label, for example a news article might be on the news and political section at the same time.
multi-class means we can treat classes as being composed of one or more classes. For example, we can have (*)

4.5 (2 points) What is the main difference between multi-class and multi-label classification?

4

SAME.

(*) classes {1,2,3} and the multi-class classificator will classify into one of $\mathcal{P}(\{1,2,3\})$ (power set of 1,2,3).
So this is still classification where the # of classes grows pretty fast.

Problem 5. Expectation-Maximization Algorithm

5.1. (3 points) What function is maximized in the EM algorithm? Provide a formula and/or a precise description.

$$\underset{\theta}{\arg\max} \left\{ E_y \left[\log p(\vec{y} \mid \mathcal{X}, \theta^{\cancel{**}}) \mid \theta^{(t)} \right] \right\}$$

2/3 We want to maximize the expectation of the likelihood of class labels assuming we have parameters $\theta^{(t)}$, i.e., an estimate of the true parameters of the mixture of distributions.

5.2. (3 points) What is the main characteristic of problems that are suitable for an application of the EM algorithm?

Mixture of distribution all with the same distribution.

For example: $\sum_{i=1}^{m} w_i P(x_j \mid \theta)$, where $P(x_j \mid \theta)$ might be

$1\frac{1}{2}/3$

Poisson, or in general any other distribution

Think more generally! Hard to maximize actual likelihood.

5.3. (2 points) In the process of estimating the parameters of a finite mixture of distributions, what is the intuition that leads to the solution we refer to as classification EM algorithm?

the intuition is to regard the labels y_i as known, i.e., to suppose we know from wich distribution each data point comes from ✓ and then maximize the respective expectation.

2

5.4. (2 points) Briefly discuss the relationship between the EM algorithm and K-means clustering?

Both of these algorithms attempt to estimate the parameters of a mixture of distributions, but K-means uses some sort of distance notion whereas EM perfoms Expectation/maximization of complete data set (label + data).

2

Problem 6. Linear Regression

6.1 (2 points) Consider a training set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$. Write an error function that is typically minimized in OLS linear regression? Do not use matrix formulation.

$$E = \sum_{i=1}^{n} \left(f(\vec{x_i}) - y_i \right)^2$$

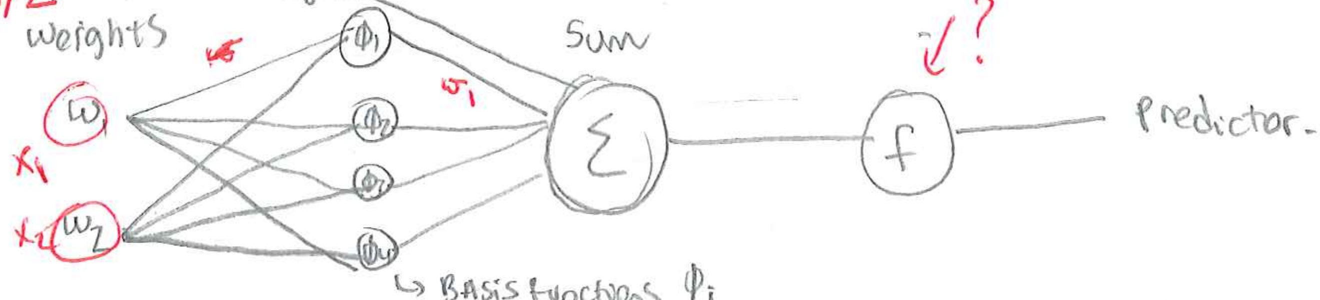
what is $f(x_i)$?

$1\frac{1}{2}/2$

Sum of squared errors

6.2 (2 points) Draw a radial basis function (RBF) network with 2-dimensional inputs and 4 basis functions. Define all variables and functions.

1/2



6.3 (3 points) Briefly describe the need for regularization in OLS regression. Then define at least one regularization method actively used in OLS regression.

Regularization is needed to avoid overfitting, i.e., to allow room for more error in our estimate hoping that the predictor will work best on test data.

two kinds of regularization method: lasso and ridge.

For example lasso $E = \sum (f(\vec{x_i}) - y_i)^2 + \lambda \sum_{i=1}^{n} |w_i|$ ← this bounds the coefficients

3

6.4 (3 points) Use matrix notation to derive a gradient descent method for find a solution to OLS regression. Assume that $\mathcal{X} = \{1\} \times \mathbb{R}^k$, and $\mathcal{Y} = \mathbb{R}$. The table below (from class) provides useful derivatives.

$\mathbf{y}$	$\partial \mathbf{y} / \partial \mathbf{x}$
$\mathbf{Ax}$	$\mathbf{A}^T$
$\mathbf{x}^T\mathbf{A}$	$\mathbf{A}$
$\mathbf{x}^T\mathbf{x}$	$2\mathbf{x}$
$\mathbf{x}^T\mathbf{Ax}$	$\mathbf{Ax} + \mathbf{A}^T\mathbf{x}$

Gradient descent is:

$$w^{(t+1)} = w^{(t)} - \eta \nabla f(X)$$

In this case:

1/3

$(\mathcal{D}|\lambda)$ $P(\lambda)$

Problem 7. Parameter Estimation

7.1. (10 points) Suppose that data set $\mathcal{D} = \{1, 0, 1, 1, 1, 0, 1, 1, 1, 0\}$ is an i.i.d. sample from a Bernoulli distribution

$$p(x|\alpha) = \alpha^x (1-\alpha)^{1-x} \qquad 0 < \alpha < 1$$

with an unknown parameter α.

a) (4 points) Calculate the log-likelihood function that $\mathcal{D}$ was generated from the Bernoulli distribution with $\alpha = 1/e$; i.e. find $\ln p(\mathcal{D}|\alpha = 1/e)$. The parameter e is the Euler number, $e \approx 2.71$. Write the final expression in as compact a form as you can.

4

Likelihood $= \arg\max_\alpha \{P(\mathcal{D}|\alpha)\} \Rightarrow P(\mathcal{D}|\alpha) = \prod_{i=1}^{10} \alpha^{x_i}(1-\alpha)^{1-x_i} = \alpha^{\sum_{i=1}^{n} x_i}(1-\alpha)^{10-\sum_{i=1}^{10} x_i}$

$ll(\alpha) = \log(\alpha^{\sum x_i}(1-\alpha)^{10-\sum x_i}) = (\sum_{i=1}^{10} x_i)\log(\alpha) + (10 - \sum_{i=1}^{10} x_i)\log(1-\alpha))$. using our data

$\sum_{i=1}^{10} x_i = 7$, $\alpha = 1/e \Rightarrow ll(1/e) = 7\log(1/e) + 3\log(1-\frac{1}{e}) = -7\log(e) + 3\log(\frac{e-1}{e})$

$\Rightarrow ll(1/e) = -7 + 3\log((e-1)\cdot e^{-1})$, $= -7 + 3[\log(e-1) - \log(e)]$ $\Rightarrow \boxed{ll(1/e) = -7 + 3[\log(e-1) - 1]}$

b) (6 points) Suppose the prior distribution for α is the uniform distribution on (0,1). Compute the Bayes estimator of α. Note that $\int_0^1 v^m(1-v)^r\,dv = \frac{m!r!}{(m+r+1)!}$.

Bayes estimator = mean of the posterior $P(\alpha|\mathcal{D})$

Note, $P(\alpha|\mathcal{D}) = \frac{P(\mathcal{D}|\alpha)\cdot p(\alpha)}{P(\mathcal{D})} = \frac{P(\mathcal{D}|\alpha)\cdot P(\alpha)}{\int_0^1 P(\mathcal{D}|\alpha)\cdot P(\alpha)\,d\alpha}$

$\Rightarrow \int_0^1 P(\mathcal{D}|\alpha)\cdot P(\alpha) = \int_0^1 \alpha^{\sum x_i}(1-\alpha)^{n-\sum x_i}\cdot 1\,d\alpha = \frac{(\sum x_i)!(n-\sum x_i)!}{(\sum x_i + n - \sum x_i + 1)!}$ ← by hint

6

$= \frac{(\sum x_i)!(n-\sum x_i)!}{(n+1)!} = C$

So, $P(\alpha|D) = \frac{\alpha^{\sum x_i}(1-\alpha)^{n-\sum x_i}}{C}$, the Bayes Estimator is

$E_B[P(\alpha|D)] = \int_0^1 \frac{\alpha^{\sum x_i}(1-\alpha)^{n-\sum x_i}}{C}\alpha\,d\alpha = \frac{(n+1)!}{(\sum x_i)!(n-\sum x_i)!}\int_0^1 \alpha^{(\sum x_i)+1}(1-\alpha)^{n-\sum x_i}\,d\alpha$

$= \frac{(n+1)!}{(\sum x_i)!\cancel{(n-\sum x_i)!}}\cdot\frac{((\sum x_i)+1)!\cancel{(n-\sum x_i)!}}{(n+2)!} = \boxed{\frac{(\sum x_i)+1}{n+2}}$

In our case $E_B = \frac{7+1}{12} = \frac{8}{12} = \boxed{\frac{2}{3}}$

Problem 8. Linear Classification

8.1 (2 points) Write the expression for $P(Y = 1|\mathbf{x})$ in logistic regression. Describe all variables.

$$P(Y=1|x) = \frac{1}{1+e^{-w^T x}}$$

$$P(Y=0) = 1 - \frac{1}{1+e^{-w^T x}}$$

8.2 (2 points) What is the main difference between a logistic regression classifier and a perceptron?

8.3 (2 points) What is the main similarity in how the logistic regression model and the perceptron are trained?

8.4 (2 points) Why are we adding a column of ones to our data matrix $\mathbf{X}$ before performing the logistic regression optimization?

8.5 (2 points) What does the Pocket algorithm minimize?

Problem 9. Classification and Regression

9.1. (2 points) How is Newton-Raphson's method used in logistic regression?

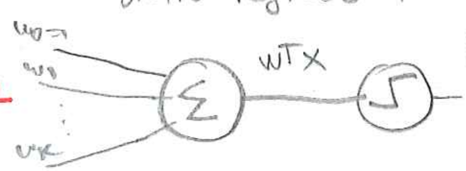
2 In logistic regression we do not have the luxury of having closed form solutions to the maximization problem of predicting class labels using the logistic function as Bernoulli probabilities. Instead, we us N-R method to derive an iterative algorithm to find an optimum solution

9.2 (2 points) What is overfitting?

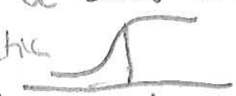
2 Overfitting is the phenomenon that occurs in learning algorithms when the algorithm performs really well on Training data but then under performs on actual test data. We say that training data was overfit or that the algorithm is not flexible enough to generalize to other data.

9.3 (2 points) Intuitively explain the notion of likelihood in the logistic regression problem.

logistic regression

2 

$\begin{cases} +1 & \text{if } w^T x \geq 0 \\ -1 & \text{if } w^T x < 0 \end{cases}$

The likelihood in this case is intuitively the probability that a point belongs to a class label. we use logistic function  => the max. likelihood here is just whether the (Δ)

(Δ) probability (after training) is ≥ 0.5 => pos. label and < 0.5 => neg. label.

9.4 (2 points) Briefly describe the relationship between the gradient descent and Newton-Raphson's method.

2 Gradient descent is a first order method that uses only first derivative (i.e., gradient). Newton-Raphson's method is a second order method that incorporates Information about the second derivative. In other words: Gradient descent is a especial case of N-R, where the second derivative is set to 1 or the identity matrix.

9.5 (2 points) If the perceptron is a non-linear function $f: \mathcal{X} \rightarrow \mathcal{Y}$ defined as

$$f(\mathbf{x}) = \begin{cases} +1 & \mathbf{w}^T\mathbf{x} \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

where $\mathbf{w}$ is a set of weights, how come we call perceptron a linear classifier? Are we wrong?

2 we call perceptron a linear classifier because the sets of weights $\vec{w}$ that the algorithm learns is such that $w^T \cdot x = 0$ defines a line that separates positives from negatives examples (in case this is possible, i.e., data is linearly separable).

Problem 10. True or False

(no explanation or justification, just write T or F in the spaces below questions)

10.1. (2 points) Consider a probability density function $p_X(x)$. The value of this function at point x_0 represents the probability that random variable X has value x_0.

T

0

10.2. (2 points) There exist 2^n distinct factorizations of a discrete joint probability distribution $P(X_1, X_2, \ldots, X_n)$ involving n random variables.

2 F

10.3. (2 points) Consider the following estimator

$$\hat{\alpha} = \arg\max_{\alpha}\{P(\alpha|\mathcal{D})P(\mathcal{D})\}$$

where $\mathcal{D}$ is some set of observations and α is a parameter. This estimator is referred to as the maximum a posteriori (MAP) estimator.

2 F

10.4 (2 points) The maximum likelihood solution to an ordinary least squares regression problem produces an unbiased estimator.

2 T

10.5 (2 points) In linear classification, the minimization of the Euclidean distance between the predictions and the class labels on the training set is guaranteed to find a global minimum.

T

0