INDIANA UNIVERSITY
CS B555 Fall 2015

Practice Final Exam

Do Not Distribute

Duration: 2 hours

Student ID: |_|_|_|_|_|_|_|

Last Name: ______________________________

First Name: ______________________________

Carefully read all of the instructions and questions. Good luck!

---

1. **Do not turn this page** until you have received the signal to start.
2. **You may use a two-page cheat sheet**. No electronic devices are allowed.
3. Please write your name on the top right corner of each page.
4. Check that the exam package has 11 pages.
5. The exam is designed for 1 hour, but you have 2 hours to complete the exam. Since you have time, attempt an answer to all parts of the problems, since the exam is worth 25%.
6. Answer all questions in the space provided; if you require more space, continue on the back of the page. For marking, I will ignore anything written on scrap paper, but please leave them with the exam.
7. Be precise, concise and give clear answers. **Use proper terminology** to get full marks.
8. If the answer is not legible, I will not be able to mark it.

Question 1. [15 MARKS]

Given dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$, the objective is to find parameters of the function

$$f(\mathbf{x}) = w_1 x_1 + w_2 x_2 + w_3 x_2 x_3$$

such that the weighted sum of square errors between the target and the prediction is minimized, with some of the samples having higher importance than other samples according to importance weights $c_1, \ldots, c_n > 0$.

Part (a) [5 MARKS]

Is f a linear function? Explain why or why not.

Part (b) [10 MARKS]

Derive an algorithm for obtaining the parameter vector $\mathbf{w} = (w_1, w_2, w_3)$. If there is a closed form solution, provide the closed-form solution; otherwise, provide an iterative, first-order gradient descent update.

Question 2. [10 MARKS]

Imagine that you want to train a classifier on a small training data set. Choose a learning algorithm and discuss two strategies to prevent or reduce overfitting. Your choice of algorithm can take into account that you want to reduce overfitting.

Question 3. [15 MARKS]

Part (a) [5 MARKS]

What is the purpose behind matrix factorization methods?

Question 4. [10 MARKS]

What is the general optimization scheme for learning matrix factorization models?

Question 5. [10 MARKS]

Suppose that you have three random variables X, Y, Z.

Part (a) [4 MARKS]

Assume X can take on any values in $[0, 1]$. It either has a probability density function or a probability mass function. Explain which it has, using an example.

Part (b) [3 MARKS]

If $P(X, Y) = P(X)P(Y)$, what does this tell us about X and Y?

Part (c) [3 MARKS]

If $P(X, Y|Z) = P(X|Z)P(Y|Z)$, what does this tell us about X, Y and Z?

Question 6. [5 MARKS]

Discuss any one form of regularization that is used to train least squares models. Why is such regularization used?

Question 7. [10 MARKS]

Part (a) [5 MARKS]

What is the main advantage of the generalized linear model (GLM) formulation of a regression problem compared to the standard OLS regression formulation?

Part (b) [5 MARKS]

Discuss why we have focused on exponential family distributions in this course.

Question 8. [10 MARKS]

Part (a) [5 MARKS]

What is the goal of MAP estimation? How does this relate to maximum likelihood estimation?

Part (b) [5 MARKS]

Intuitively explain the notion of likelihood in the logistic regression problem.

Question 9. [15 MARKS]

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

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

with an unknown parameter α.

Part (a) [5 MARKS]

Calculate the log-likelihood function that $\mathcal{D}$ was generated from a 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.

Part (b) [10 MARKS]

Suppose the prior distribution for α is the uniform distribution on $[0,1]$. Compute $p(\mathcal{D})$. Note that

$$\int_0^1 v^m(1-v)^r dv = \frac{m!r!}{(m+r+1)!}.$$

# 1	# 2	# 3	# 4	# 5	# 6	# 7	# 8	# 9	Total
/15	/10	/15	/10	/10	/5	/10	/10	/15	/100