UNIVERSITY OF ALBERTA
CMPUT 466/566 Fall 2018

Practice Midterm Exam

Do Not Distribute

Duration: 80 minutes

Last Name: Lv

First Name: Zhonghao

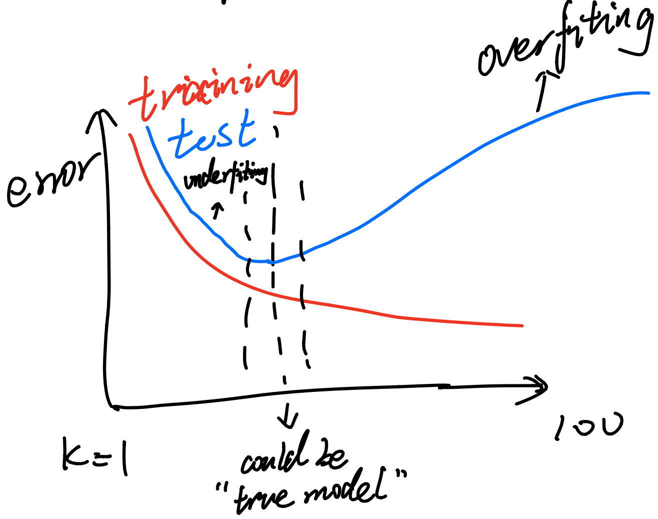
Carefully read all of the instructions and questions. Good luck!

1. **Do not turn this page** until you have received the signal to start.
2. **You may use a two-page cheat sheet**, which is four pages front-and-back. 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 8 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 35%.
6. Answer all questions in the space provided; if you require more space, you can get a blank piece of paper from the front, write the answer on that with the question number clearly labeled and hand it in with your exam.
7. Be precise, concise and give clear answers.
8. If the answer is not legible, I will not be able to mark it.

Question 1. [20 MARKS]

Imagine you transform the input observations to higher-order polynomials, before using linear regression. You consider all polynomials of orders $k = 1, 2, ..., 100$. You are given a training set, with a separate test set. How might you determine which models are overfitting or underfitting?

overfitting / underfitting.



example:

$$F(w) = w + w x_1 + w x_2 + w x_1 x_2 + w x_1^2 + w x_2^2$$

Name: ____________________

Question 2. [10 MARKS]

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

Part (a) [4 MARKS]

Assume X can take any values in $[0,1]$ (i.e., its outcome space is $[0,1]$). It either has a probability density function (pdf) or a probability mass function (pmf). Explain which it has, using an example of a possible pdf or pmf for X.

pdf, outcome spaces is continous.

example: uniform distribution. (clib gaussian, ...)

$$p(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b. \\ 0, & \text{otherwise} \end{cases} \Big/ \; p(x) = \frac{1}{b-a} \; x \in [a,b]$$

Part (b) [3 MARKS]

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

X, Y Independent. 补充说明

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?

conditionally independent

Question 3. [5 MARKS]

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

L_1 / L_2 regularization.

L_2: (Rither ~)

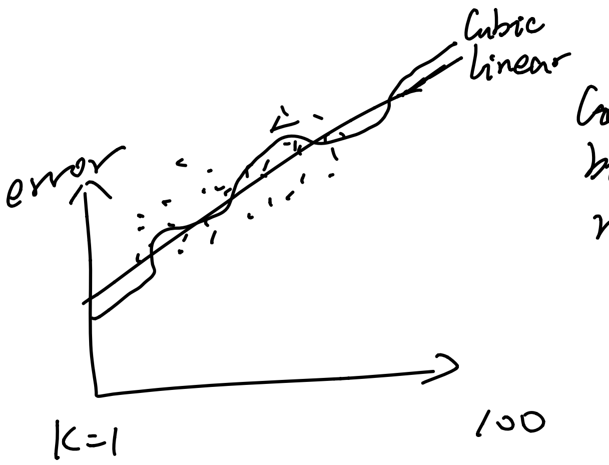
→ prevent overfitting.

L_1: feature selection, (maybe you have a lot features).

(Lasso)

Question 4. [20 MARKS]

Let us assume a setting where the true model is linear, i.e., $Y = w_0 + \sum_{j=1}^{d} w_i X_i + \epsilon$ for weights $w_j \in \mathbb{R}$, random variables X_j and $\epsilon \sim \mathcal{N}(0, 1)$. Imagine you get a dataset with $n = 100$ samples, and train one model with linear regression and one model with cubic regression—linear regression, where you first expand the features into all the polynomial terms in a cubic polynomial. Which model do you think will obtain lower training error—or will they perform the same—and why?



Cubic. may have less tring error, but it overfits: when using new training data, Cubic have greater error.

Question 5. [20 MARKS] GLM

When we talked about optimal regression models, we talked about minimized expected cost under a squared error

$$\min_{f \in \mathcal{F}} \int_{\mathcal{X} \times \mathcal{Y}} p(\mathbf{x}, y)(f(\mathbf{x}) - y)^2 d\mathbf{x} dy$$

for our hypothesis space $\mathcal{F}$: the space of functions you are restricted to, such as linear functions. We found that the optimal solution, assuming $\mathcal{F}$ contains all functions, was $f(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$. In GLMs—like linear regression, logistic regression and Poisson regression—did we learn $f(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$? If not, say why not. If yes, say how we learned that f.

Question 6. [30 MARKS]

Gaussian.
Possian
Gamma
Bi 伯努利

The univariate Kumaraswamy distribution has the pdf,

$$p(x|a,b) = abx^{a-1}(1-x^a)^{b-1}$$

for parameters $a, b > 0$, where $x \in [0,1]$. Suppose we're given a set of observations, $D = \{x_1, x_2, ..., x_n\}$, with $x_i \in \mathbb{R}_+$, sampled from a Kumaraswamy distribution with unknown parameters a_0, b_0.

Part (a) [10 MARKS]

Write down a formula for $\ell(\beta) = \log p(D|a,b)$.

$X \sim \text{Normal}(\mu, \sigma)$, $D = \{X_i\}_{i=1,2,\cdots n}$.

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$P(D|\mu) = \prod_{i=1}^{n} P(X_i|\mu)$$

Independence.

$$\ln P(D|\mu) = \ln\left(\prod_{i=1}^{n} P(X_i|\mu)\right)$$

$$= \sum_{i=1}^{n} \ln(P(X_i|\mu))$$

$$= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

$$= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\frac{\partial}{\partial\mu} \ln P(D|\mu) = \sum_{i=1}^{n} + \frac{\cancel{2}(x-\mu)}{\cancel{2}\sigma^2} = 0$$

$$= D\sum_{i=1}^{n}(X_i - \mu) = 0$$

Part (b) [20 MARKS]

$$\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu = 0 \Rightarrow \mu = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Explain how you would obtain the maximum likelihood estimate of a and b given D. Include your derivations.

MAP

$$P(D|\mu)\,P(\mu) = \left(\prod_{i=1}^{n} P(X_i|\mu)\right) \cdot P(\mu)$$

MLE

1. $P(D|\mu)$ $P(\mu)$
2. $\ln P(D|\mu) + \ln P(\mu)$
3. $\frac{\partial}{\partial\mu} \ln P(D|\mu) + \frac{\partial}{\partial v} \ln(P(\mu))$
4. solve for μ.

Bonus (Mandatory for students in 566). [20 MARKS]

A typical goal behind data normalization is to make all the features of the same scale. For example, the features are rescaled to be zero-mean, with unit variance, by taking the data matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and centering and normalizing each column: $\mathbf{X}_{ij} = \frac{\mathbf{X}_{ij}-\mu_j}{\sigma_j}$ where $\mu_j = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{ij}$ and $\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_{ij}-\mu_j)^2$. Why might this be important for gradient descent? Hint: Consider a setting where you have two features, where the first has range $[0, 0.01]$ and the other $[0, 1000]$. Think about the stochastic gradient descent update for linear regression, and what issues might arise.

$$X \sim \begin{bmatrix} [0, 0.01] \\ [0, 1000] \end{bmatrix} \quad X_i = \begin{bmatrix} 0.001 \\ 750 \end{bmatrix}, y_i \quad , e_i = (y_i - \vec{w}^T \vec{x}_i)^2$$

$$\vec{W} = \vec{W} - \alpha\sqrt{e_i}\,\vec{X}_i$$

$$= \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - \alpha\sqrt{e_i}\begin{bmatrix} 0.001 \\ 750 \end{bmatrix}$$

✓

we want to bring the variance to the same scale, to easy set stepsize.

# 1	# 2	# 3	# 4	# 5	# 6	Total
/20	/10	/5	/20	/20	/30	/105