/ Daniel Santos

$\frac{1}{\sqrt{2\pi}\,\sigma} \cdot e$

75/100

COMP 4220: Machine Learning, Fall 2018

Exam 1

Date: October 15, 2018

20/30

1. Let X and Y be two random variables, β a constant, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ be a Gaussian random variable with zero mean and variance σ^2. We assume that $Y = \beta X + \epsilon$, and that ϵ is independent of X.

   (a) Show that given $X = x$, the distribution of Y is $\mathcal{N}(\beta x, \sigma^2)$.

   (b) Let $\{(X^{(i)}, Y^{(i)}), i = 1, \ldots, n\}$ be n independent samples from the model above. Show that the maximum likelihood estimation of β has the following:

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left(Y^{(i)} - \beta X^{(i)}\right)^2$$

   (c) Show that solution of the above problem is:

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} Y^{(i)} X^{(i)}}{\sum_{i=1}^{n} \left(X^{(i)}\right)^2}$$

a) $P(Y \mid X{=}x, \beta) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}(Y-\beta x)^2}$ ← This migth be wrong. I forgot the Gaussian Distribution formula in Cheat Sheet.

10/10

b) The above will yield something in the form of $P(Y \mid X{=}x, \beta) \sim N(\beta x, \sigma^2)$, I didnt have the Gaussian distribution formula. My closest guess is the following:

$$\beta = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}(y^{(i)} - \beta x^{(i)})^2}$$

$$\log \beta = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}(y^{(i)} - \beta x^{(i)})^2}\right) = \cancel{\sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)} + \sum_{i=1}^{n} \log\left(e^{-\frac{1}{2}(y^{(i)} - \beta x^{(i)})^2}\right)$$

Drop this because is constant.

10/10

$$\hat{\beta} = \arg\max_{\beta} -\frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - \beta x^{(i)}\right)^2$$

$$\hat{\beta} = \arg\min \frac{1}{2} \sum_{i=1}^{n} \left(y^{(i)} - \beta x^{(i)}\right)^2$$

c) $\hat{\beta} = \frac{P(Y \mid X, \beta)\, P(x)}{P(Y)}$

0/10

10/30

2. Let X and Y be two random variables, and $Y = \beta X + \epsilon$, β is a constant, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Given n independent sample points $\left(X^{(1)}, Y^{(1)}\right), \ldots, \left(X^{(n)}, Y^{(n)}\right)$, instead of ordinary least squares, here we estimate β with "ridge regression", by solving the following problem:

$$\widehat{\beta} = \arg\min_{\beta} \frac{1}{2}\left(\sum_{i=1}^{n} \left(Y^{(i)} - \beta X^{(i)}\right)^2 + \lambda\beta^2\right)$$

where $\lambda \geq 0$ is a tuning parameter.

(a) Give a solution in explicit form for $\widehat{\beta}$.

(b) When λ goes to infinity, how does $\widehat{\beta}$ change? Give a brief explanation. 0/10

a)

$$\nabla_\beta \frac{1}{2}\left(\sum_{i=1}^{n}(y^{(i)} - \beta x^{(i)})^2 + \lambda\beta^2\right)$$

$$\nabla_\beta \frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - \beta x^{(i)})^2 + \nabla_\beta \lambda\beta^2$$

$$\nabla_\beta \frac{1}{2}\left(\beta x^{(i)} - y^{(i)}\right) + \nabla_\beta(\lambda\beta^2)$$ 10/20

$$0 = -\beta^T y + \beta^T\beta X + 2\lambda\beta$$

$$X = (\beta^T\beta + 2\lambda I)^{-1}\beta^T y$$

15/30

3. Suppose you are given the following classification task: predict the target $y \in \{0, 1\}$ given two real valued features x_1 and x_2. After some training, you learn the following decision rule:

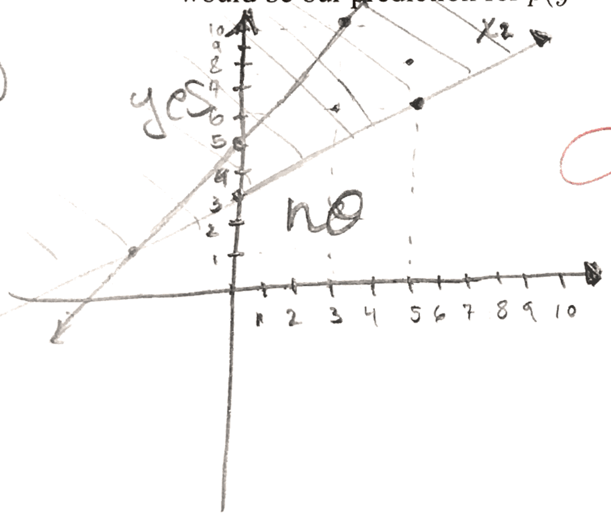
$$\text{“}y = 1 \text{ if } \underbrace{\overset{3}{w_1}x_1 + \overset{5}{w_2}x_2 + \overset{-15}{w_0}} \geq 0 \text{ and } y = 0 \text{ otherwise”}$$

where $w_1 = 3, w_2 = 5, w_0 = -15$.

(a) Plot the decision boundary and label the region where we would predict $y = 1$ and $y = 0$.

(b) Suppose that we learned the above weights using logistic regression. Using this model, what would be our prediction for $p(y = 1|x_1, x_2)$?

a)



$3X_1 + 5X_2 - 15 \geq 0$

$3X_1 + 5X_2 \geq 15$

$X_1 \geq \frac{5}{3}X_2 + 5$

$X_2 \geq \frac{3}{5}X_1 + 3$

10/20

b) $\sigma(w_0 + w_1x_1 + w_2x_2) = \dfrac{1}{1 - \exp(-w_0 - w_1x_1 - w_2x_2)}$?

5/10

30
30

4. A set of data points is generated by the following process: $Y = w_0 + w_1 X + w_2 X^2 + w_3 X^3 + w_4 X^4 + \epsilon$, where X is a real-valued random variable and ϵ is a Gaussian noise variable. You use two models to fit the data:

$E[\hat{Y}] - \cancel{Y}$

Model 1: $Y = a_0 + a_1 X + \epsilon$
Model 2: $Y = a_0 + a_1 X + \ldots + a_9 X^9 + \epsilon$

(a) Model 1, when compared to Model 2 using a fixed number of training examples, has a *bias* which is:

- ◇ Lower
- ◇ Higher
- ◇ The Same

(b) Model 1, when compared to Model 2 using a fixed number of training examples, has a *variance* which is:

- ◇ Lower
- ◇ Higher
- ◇ The Same

(c) Given 12 training examples, which model is more likely to *overfit* the data?

- ◇ Model 1
- ◇ Model 2