$\frac{227}{220}$

CS 4900/5900 Exam (Oct 12, 2017) Name: BHISHAN POUDEL

Problem 1 (30 points)

30

Suppose there are two cookie bowls, one red and one blue. The red bowl has 10 chocolate chip and 30 plain cookies, while the blue bowl has 20 of each. Hui picks a bowl at random, and then picks a cookie at random. We may assume there is no reason to believe Hui treats one bowl differently from another, likewise for the cookies. The cookie turns out to be a plain one. How probable is it that Hui picked it out of the red bowl? Explain your reasoning.

Red	Blue	Total
10C	20C	30 chocolates
30P	20P	50 plain cookies
40 red	40 blue	80 things

$p(r) = 0.5 = \frac{1}{2}$ $\quad p(b) = 0.5 = \frac{1}{2}$ choosing red & blue bowl has equal probabilities

$P(p|r) = \frac{30}{40}$ $\quad p(p|b) = \frac{20}{40} = \frac{1}{2}$

prob of choosing plain cookie from red bowl

$p(r|p) = ?$

prob that chosen plain cookie came from red bowl

using Bayes's theorem,

$$p(r|p) = \frac{p(p|r)\,p(r)}{p(p)} = \frac{p(p|r)\cdot p(r)}{p(p|r)\cdot p(r) + p(p|b)\cdot p(b)}$$

$$= \frac{\frac{3}{4}\cdot\frac{1}{2}}{\frac{3}{4}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{2}{2}} = \frac{3/8}{(3+2)/8} = \frac{3/8}{5/8} = \frac{3}{5}$$

$\therefore \boxed{p(r|p) = \frac{3}{5}}$ Ans

27

Problem 2 (30 points)

X is a random variable that is normally distributed $N(\mu, \sigma^2)$ i.e. the probability density function is:

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

The parameters μ and σ of the distribution are not known, however we observe a sequence $x_1, x_2, ..., x_n$ of n independent samples of X. Use the Maximum Likelihood estimation principle to estimate the mean μ of the distribution from the n samples.

prob dist fn $p(x_n) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$

likelihood function $L(\mu) = \prod_{n=1}^{N} p(x_n) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$ ✓

maximum likelihood estimate of μ is $\vec{\mu} = \underset{\mu}{\arg\max}\ L(\mu)$

$\vec{\mu} = \underset{\mu}{\arg\min}\ (-)\ln L(\mu)$ ✓

$= -\underset{\mu}{\arg\min}\ \ln \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}}$ ✓

then,

$0 = \frac{\partial}{\partial\mu}\left[-\sum_n \frac{(x_n-\mu)^2}{2\sigma^2} - N\ln\sqrt{2\pi\sigma^2}\right]$

$0 = +\sum_n \frac{\cancel{2}(x_n-\mu)}{\cancel{2}\sigma^2}$ ✓

$= \sum_n x_n - \sum_n \mu$

$0 = \sum_{n=1}^{N} x_n - N\vec{\mu}$

$$\boxed{\vec{\mu} = \frac{\sum_{n=1}^{N} x_n}{N}}$$

Ans

$\log x \cdot y = \log x + \log y$

✓

$\ln \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/2\sigma^2}$

$= \sum_{n=1}^{N} \ln \frac{e^{-(x_n-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$ $\log \frac{a}{b} = \log a - \log b$

$\sum_{n=1}^{N} \ln e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} - \sum_{n=1}^{N} \ln\sqrt{2\pi\sigma^2}$

$= -\sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2} - N \cdot \ln\sqrt{2\pi\sigma^2}$

X degree of polynomial

$1 \quad x_1 \quad (x_1)^2 \cdots (x_1)^M$

$1 \quad x_2 \quad (x_2)^2 \cdots (x_2)^M$

$1 \quad x_N \quad (x_N)^2 \cdots (x_N)^M$

$x_1 \; t_1 \quad h_1 = w^T x_1$

$x_2 \; t_2 \quad h_2 = w^T x_2$

$x_N \; t_N \quad h_N = w^T x_N$

Problem 3 (40 points)

40

Let $D = \{(x_1, t_1), (x_2, t_2), ..., (x_N, t_N)\}$ be a training dataset for learning a polynomial regression function $y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j$.

$SSE = \sum_n (h_n - t_n)^2$

$h = Xw^T$

(a) SSE Write the formula for the Sum-of-Squares error function:

X shape $= N, m+1$

w shape $= 1, m+1$

(I choose row vector)

$$E(\mathbf{w}) = (h - t)^2 = \sum_{n=1}^{N} (h_n - t_n)^2$$

matrix

where $h_n = x_n w^T$ $\quad w = [w_0, w_1, \dots, w_M]$

t shape $= N, 1$ column vector

(b) Write the formulas for the Root Mean Square Error and the Mean Absolute Error of the model $y(x, \mathbf{w})$ on the dataset D:

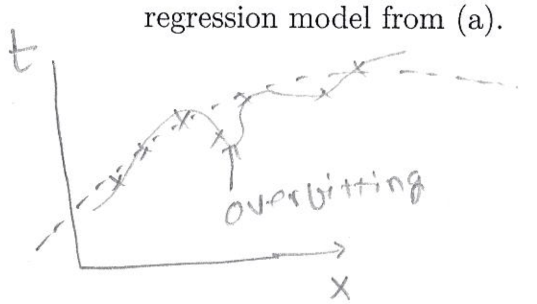
$h = X w^T = (N, m+1) @ (m+1, 1)$

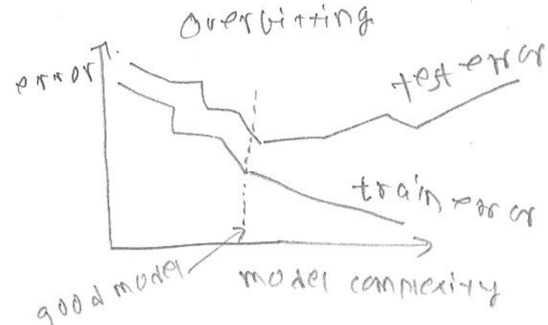
h shape $= N, 1 =$ same shape of t

$$RMSE(D) = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (h_n - t_n)^2}$$

$$MAE(D) = \frac{|h - t|}{N} = \sum_{n=1}^{N} \frac{|h_n - t_n|}{N}$$

(c) Define overfitting. Describe two methods that can be used to reduce overfitting in the regression model from (a).





Solution of overfitting

1) increase # of training example

2) reduce number of features

when increasing # of features if train error goes very small but test error goes large then we call the model is overfitting the data

(d) What is Occam's Razor? How can be Occam's Razor implemented in a linear regression model?

Occam's Razor : If number of different model fits the same dataset equally good we should choose the simplest model.

Usage : In linear regression if we have N data points, a polynomial of degree N will perfectly fit all the data points in train set, but it may not give good result in test set since the model catches all the noise in trainset. we use Occam's razor and use lower degree polynomial

e.g. cubic polynomial to fit the data that does good on both

32

Problem 4 (40 points)

Consider a dataset that contains the 4 examples below i.e., the truth table of the logical XOR function. (a) Show the formula used to compute the output of a logistic regression model on feature vector **x**, given parameters vector **w**. (b) What is the criterion used to classify example **x** as positive? Prove that no logistic regression model can perfectly classify this dataset. Do not forget the bias feature $x_0 = 1$. (c)

missing. ok.

x_1	x_2	t
0	0	0
0	1	1
1	0	1
1	1	0

Hint: Prove that there cannot be a vector of parameters **w** such that $P(t = 1|\mathbf{x}, \mathbf{w}) \geq 0.5$ for all examples **x** that are positive, and $P(t = 1|\mathbf{x}, \mathbf{w}) < 0.5$ for all examples **x** that are negative.

(a) logistic regression: $P(C_1|x) = \sigma(w^T x) = \dfrac{1}{1+e^{-w^T x}}$

prob of data belong to class C_1

irrelevant

$P(C_2|x) = 1-\sigma = \dfrac{e^{-w^T x}}{1+e^{-w^T x}}$

the formula to compute cost for binary logistic regression is

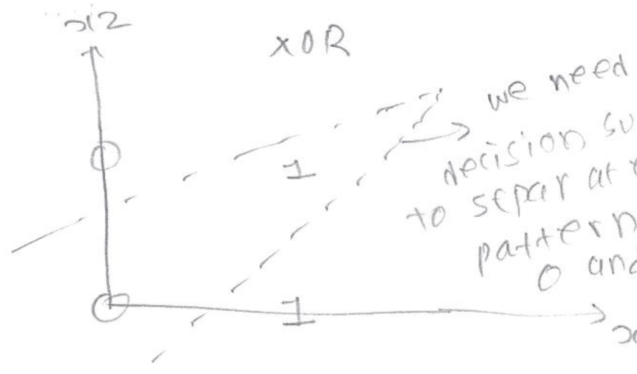
$E = -\sum_n \left[t_n \ln h_n + (1-t_n) \ln (1-h_n) \right]$ where $h = w^T x$

(b) classification criteria: if $\dfrac{1}{1+e^{-w^T x}} \geq \dfrac{1}{2}$ data belongs to class 0

if $\dfrac{1}{1+e^{-w^T x}} < \dfrac{1}{2}$ data belongs to class 1.

(c) Truth table of XOR

x_1	x_2	XOR
0	0	0
0	1	1
1	0	1
1	1	0



we need to decision surface to separate patterns 0 and 1 from the figure we see that XOR logic is not linearly separable.

X			t
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	0

The examples in dataset X will be linearly separable

if $w_0 + \sum_i w_i x_i \geq 0$ if $t_i = 1$

and $w_0 + \sum_i w_i x_i < 0$ if $t_i = 0$?

Now,

$w_0 < 0$

$w_0 + w_2 \geq 0$

$w_0 + w_1 \geq 0$

$w_0 + w_1 + w_2 < 0$

$\longrightarrow w_0 + (w_0 + w_1 + w_2) \geq 0$

but $w_0 < 0$ and $w_0 + w_1 + w_2 < 0$
so this contradicts our assumption of linear separability.

$\therefore$ Dataset X is NOT linearly separable.

this is not a proof

20

Problem 5 (40 points)

Consider the training dataset $\mathcal{D}$ below, where a_1 and a_2 are two discrete attributes and l is the label.

$\mathcal{D}$	$a_1(\mathbf{x})$	$a_2(\mathbf{x})$	$l(\mathbf{x})$
$\mathbf{x}_1$	red 0	fish 001	cute +1
$\mathbf{x}_2$	blue 1	fish 001	ugly +1
$\mathbf{x}_3$	red 0	fly 010	ugly +1
$\mathbf{x}_4$	red 0	frog 100	ugly −1
$\mathbf{x}_5$	blue 1	fly 010	cute −1

(a) Explain how you would create an equivalent representation for the 5 training examples as vectors of features, where each feature takes a numeric value. Show the new dataset as a set of 5 feature vectors.

(b) Are the 5 examples linearly separable i.e. is there a vector $\mathbf{w}$ and a threshold τ such that $\mathbf{w}^T\mathbf{x} \geq \tau$ if and only if the example $\mathbf{x}$ is cute?

(c) Identify one training example $\mathbf{x}_i$ such that, when we eliminate $\mathbf{x}_i$ from the training set, the remaining 4 examples are linearly separable. Justify.

(a)

	w0	w1	w2	w3	w4	Target
x1	1	0	0	0	1	1
x2	1	1	0	0	1	0
x3	1	0	0	1	0	0
x4	1	0	1	0	0	0
x5	1	1	0	1	0	0

bias column

Fish = 001
fly = 010
frog = 100

0

(b)

$w_0 + w_4 \geq 0$

$w_0 + w_1 + w_4 < 0 \Rightarrow \underbrace{w_1}_{\text{negative}} + \underbrace{(w_0 + w_4)}_{\text{non-negative}} < 0$

$w_0 + w_3 < 0$

$w_0 + w_2 < 0$

$w_0 + w_1 + w_3 < 0 \Rightarrow w_1 < 0$

this is not a proof.

(c)

Problem 6 (40 points)

20

Consider the training and test datasets shown below, where each example has 3 features:

		Train			Test	Train min	Train range	Train mean	Train standard deviation
		$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$				
floor	ϕ_1	0	1	2	-1	0	2	1	1
bed	ϕ_2	1	2	3	2	1	2	2	1
age	ϕ_3	2	3	4	5	2	2	3	1

Table 1: Training and Test datasets.

$\frac{5-2}{2} > 1 \Rightarrow 1$

1. Scale the features in the dataset from Table 1 to be between [0, 1]. Show the resulting dataset in a new table, using the same format as Table 1.

2. Standardize the features in the dataset from Table 1. Show the resulting dataset in a new table, using the same format as Table 1. For the standardized values, you do not have to compute the final numbers, you can leave them in fractional form.

① min max scaling $\quad \tilde{x} = \frac{x - x_{min}}{x_{max} - x_{min}}$

feature	Train x_1	x_2	x_3	Test
ϕ_1	$\frac{0-0}{2}=0$	$\frac{1-0}{2}=0.5$	$\frac{2-0}{2}=1$	$\frac{-1-0}{2}=-0.5$
ϕ_2	$\frac{1-1}{2}=0$	$\frac{2-1}{2}=0.5$	$\frac{3-1}{2}=1$	$\frac{2-1}{2}=0.5$
ϕ_3	$\frac{2-2}{2}=0$	$\frac{3-2}{2}=0.5$	$\frac{4-2}{2}=1$	$\frac{5-2}{2}=1.5>1$ so take 1 (since we are normalizing between 0 and 1)

2

② standard normalization (Z-score normalization) $\quad \tilde{x} = \frac{x-\mu}{\sigma} \qquad \sigma = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N-1}}$

	Train x_1	x_2	x_3	Test
ϕ_1	-1	0	1	$-\cancel{1}-1=-2$ take -1
ϕ_2	-1	0	1	0
ϕ_3	-1	0	1	2

0

Rough

$\sigma_1^2 = \frac{(0-1)^2+(1-1)^2+(2-1)^2}{3-1} = \frac{1+0+1}{2} = \frac{2}{2} = 1$

$\sigma_2^2 = 1 = \sigma_3^2$

vanilla gradient descent

$v^{\tau+1} = \eta \nabla J(w^{\tau})$

$w^{\tau+1} = w^{\tau} - v^{\tau+1}$

Bonus 1 (15 points)

Write down the objective function for Lasso. Explain all notation used.

58

$J = \frac{1}{2N} \sum_{n=1}^{N} (h_n - t_n)^2 + \frac{\lambda}{2} \sum_{j=1}^{M} |w_j|$ ✓

we donot regularize w_0

λ = penalty parameter

$h = w^T x$ = hypothesis

w = weight vector

Bonus 2 (15 points)

Write down the gradient update for gradient descent with momentum. Explain all notation used.

$v^{\tau+1} = \gamma v^{\tau} + \cancel{\alpha}\ \eta \nabla J(w^{\tau})$

$w^{\tau+1} = w^{\tau} - v^{\tau+1}$

τ = iteration

η = learning rate

Bonus 3 (15 points)

Write down the gradient update for Nesterov accelerated gradient. Explain all notation used.

$v^{\tau+1} = \gamma v^{\tau} + \eta \nabla J(w^{\tau} - \gamma v^{\tau})$

$w^{\tau+1} = w^{\tau} - v^{\tau+1}$ ✓

τ = time stamp or iteration number

η = learning rate

γ = momentum hyperparameter ($\gamma \approx 0.9$ usually)

Bonus 4 (15 points)

What is the value computed by the following statement in Numpy? Explain all intermediate computations and show all intermediate results.

[0 2 4 6 8 10]

```
np.arange(0, 12, 2).reshape(3,2).T.ravel().reshape(2,3).dot([-1, 0, 1])
```

$a = [0, 2, 4, 6, 8, 10]$

$b = a.\text{reshape}(3,2) = \begin{bmatrix} 0 & 2 \\ 4 & 6 \\ 6 & 8 \end{bmatrix}_{3,2}$

$b.T = \begin{bmatrix} 0 & 4 & 6 \\ 2 & 6 & 8 \end{bmatrix}_{2,3}$

$c = b.\text{ravel}() = [0\ 4\ 6\ 2\ 6\ 8]$

$d = c.\text{reshape}(2,3) = \begin{bmatrix} 0 & 4 & 6 \\ 2 & 6 & 8 \end{bmatrix}$

$d.\text{dot}([-1, 0, 1]) = \begin{bmatrix} 0 & 4 & 6 \\ 2 & 6 & 8 \end{bmatrix}_{2\times3} \cdot \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}_{3\times1}$

$r = \begin{bmatrix} 0\times-1 + 0 + 6 \\ \ldots \end{bmatrix}_{2\times1}$

$\begin{bmatrix} 8 \\ 8 \end{bmatrix}$

np. arange (0, 12, 2) = [0, 2, 4, 6, 8, 10]

. reshape (3, 2) $= \begin{bmatrix} 0 & 2 \\ 4 & 6 \\ 8 & 10 \end{bmatrix}$ shape = 3, 2

. T $= \begin{bmatrix} 0 & 4 & 8 \\ 2 & 6 & 10 \end{bmatrix}$ shape = 2, 3

. ravel ( ) = [0 4 8 2 6 10] shape = (1,)

. reshape (2, 3) $= \begin{bmatrix} 0 & 4 & 8 \\ 2 & 6 & 10 \end{bmatrix}$ shape = 2, 3

. dot ([-1, 0, 1]) $= \begin{bmatrix} 0 & 4 & 8 \\ 2 & 6 & 10 \end{bmatrix}_{(2,3)} \cdot \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}_{3,1}$

$= \begin{bmatrix} 8 \\ -2 + 10 \end{bmatrix}_{2,1}$

$= \begin{bmatrix} 8 \\ 8 \end{bmatrix}$ Ans