COL 341 Minor 1

Chinmay Mittal

TOTAL POINTS

65.5 / 70

QUESTION 1

Generative - Linear Regression 25 pts

1.1 Q1a **8 / 10**

✓ **+ 10 pts** *Correct*

+ 0 pts Incorrect

- 2 *Point adjustment*

Had to mention about independence assumption in step P(D/w) = Product(P(x_i,y_i/w).

1.2 Q1b **10 / 10**

✓ **+ 10 pts** *Correct*

+ 0 pts Click here to replace this description.

1.3 Q1c **5 / 5**

✓ **+ 5 pts** *Correct*

+ 0 pts incorrect/not attempted

- 0.5 pts others

+ 1 pts did not answer the question

- 2 pts Click here to replace this description.

Note that the answer is partially correct.

You should have proven the hessian to be positive semidefinite.

2) If the hessian matrix is diagonal, then diagonal elements of the hessian >=0 iff it is positive semidefinite i.e., all its eigenvalues are >=0. This holds true because for a diagonal matrix, the diagonal elements are its eigenvalues.

3) However, in this case, the hessian matrix is not diagonal. So, you have to show that it is positive semidefinite in some other way.

QUESTION 2

GNB - Logistic 20 pts

2.1 Q2a **4.5 / 5**

✓ **+ 5 pts** *Correct*

+ 3 pts Partially Correct

+ 0 pts Incorrect

- 0.5 *Point adjustment*

The expression for Gaussian is wrong in the denominator.

2.2 Q2b **10 / 10**

Click here to replace this description.

- 10 pts Incorrect/Unattempted

✓ **- 0 pts** *correct*

2.3 Q2c **5 / 5**

+ 4.5 pts correct

+ **0 pts** incorrect/not attempted

+ **3.5 pts** partially correct

✓ **+ 5 pts** *correct*

+ **1 pts** Click here to replace this description.

QUESTION 3

3 VC Dimension **8 / 10**

+ **10 pts** Correct

✓ **+ 8 pts** *Mostly Correct - minor bug*

+ **4 pts** Major bugs/ Significant explanation missing

+ **2 pts** Incorrect - some reasonable attempt

+ **0 pts** Not Attempted/ Incorrect

VC-d is infinite.

QUESTION 4

4 PLA **15 / 15**

✓ **+ 15 pts** *Correct*

+ **0 pts** Incorrect or not attempted

+ **0 pts** Correct but incomplete arguments (partial marks awarded via 'Point Adjustment')

Student Name: CHINMAY MITTAL	Entry Number: 2020CS10336.

Department of Computer Science and Engineering
Indian Institute of Technology Delhi

COL341: Fundamentals of Machine Learning

Minor 1

Time: 60 minutes | Maximum Marks: 70 | Number of Questions: 4

Instructions: Please attempt all questions. If you feel any question/statement is ambiguous, please write your assumptions clearly, and then answer as per your assumptions.

Question	1(a)	1(b)	1(c)	2(a)	2(b)	2(c)	3	4	Total
Max Marks	10	10	5	5	10	5	10	15	70
Earned Marks									

1. In the class we discussed loss/error function for linear regression in a discriminative style. In this question we will try to derive the expression in a generative style. Consider the dataset $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}$. Assume that the value of y is observed after corruption with a Gaussian noise, $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In other words, $\mathbb{P}(y \mid \mathbf{x}, \mathbf{w}) = \mathcal{N}(y \mid \mathbf{w}^\top \mathbf{x}, \sigma^2)$

   (a) [10 marks] The maximum likelihood estimate of $\mathbf{w}$ is given by: $\mathbf{w}^*_{\text{MLE}} = \arg\max_{\mathbf{w}} \mathbb{P}(\mathcal{D} \mid \mathbf{w})$. Show that, $\mathbf{w}^*_{\text{MLE}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \left(y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)}\right)^2$

   (b) [8+2 marks] One problem with maximum likelihood estimation is that it can result in over-fitting when the number of samples are small. In the assignment we suggested you to use ridge regression to ameliorate the problem. In this question we will show that ridge regression is equivalent to MAP estimation of $\mathbf{w}$. Assuming Gaussian prior over the weight vector $\mathbf{w}$, i.e., assuming $\mathbb{P}(\mathbf{w}) = \prod_j \mathcal{N}(w_j \mid 0, \tau^2)$, show that, $\mathbf{w}^*_{\text{MAP}} = \arg\min_{\mathbf{w}} \left[\lambda \|\mathbf{w}\|_2^2 + \sum_{i=1}^{N} \left(y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)}\right)^2\right]$. What is the value of λ?

   (c) [5 marks] Notice that the error function for the ridge regression doesn't have an implicit solution, unlike linear regression. Hence, we need to solve it through gradient descent. Can you prove that the function is convex?

$$P(D|\omega) = \prod_{i=1}^{d} P(d_i|\omega) = \prod_{i=1}^{d} P(y_i|\omega, x_i)$$

$$= \prod_{i=1}^{d} \mathcal{N}(y_i \mid \omega^T x_i, \sigma^2)$$

$$W^*_{MLE} = \text{argmax}_\omega \left(\prod_{i=1}^{d} \mathcal{N}(y_i \mid \omega^T x_i, \sigma^2)\right)$$

$$= \text{argmin}_\omega \left(-\log\left(\prod_{i=1}^{d} \mathcal{N}(y_i \mid \omega^T x_i, \sigma^2)\right)\right)$$

$$w^*_{MLE} = \underset{w}{\text{argmin}} \left(- \sum_{i=1}^{d} \log \left(y_i \mid w^T x_i, \sigma^2 \right) \right)$$

$$= \underset{w}{\text{argmin}} \left(- \sum_{i=1}^{d} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(\frac{-(y_i - w^T x_i)^2}{\sigma^2} \right) \right) \right)$$

$$= \underset{w}{\text{argmin}} \left(\sum_{i=1}^{d} \underbrace{\frac{(y_i - w^T x_i)^2}{\sigma^2}}_{\text{constant wrt } \sigma} + \underbrace{\text{constant}}_{\text{wrt } w} \right)$$

$$= \underset{w}{\text{argmin}} \left(\sum_{i=1}^{d} (y_i - w^T x_i)^2 \right)$$

1b). $w^*_{MAP} = \underset{w}{\text{argmax}}\ p(D \mid w)\, p(w)$ (MAP estimate).

$$= \underset{w}{\text{argmin}} \left[-\log \left(p(D \mid w) \right) p(w) \right] = \underset{w}{\text{argmin}} \left[-\log(p(w)) - \log p(D \mid w) \right]$$

$$= \underset{w}{\text{argmin}} \left[-\log \left(\prod_j N(w_j \mid 0, \tau^2) \right) + \underbrace{\sum_{i=1}^{d} \frac{(y_i - w^T x_i)^2}{\sigma^2} + K}_{\text{from 1a}} \right]$$

$$= \underset{w}{\text{argmin}} \left[- \sum_{j=1}^{d} \log \left(\exp \left(- \frac{(w_j)^2}{\tau^2} \right) \right) + \sum_{i=1}^{d} \frac{(y_i - w^T x_i)^2}{\sigma^2} + K \right]$$

$$= \underset{w}{\text{argmin}} \left[\sum_{j=1}^{d} \frac{w_j^2}{\tau^2} + \sum_{i=1}^{d} \frac{(y_i - w^T x_i)^2}{\sigma^2} \right]$$

$$= \underset{w}{\text{argmin}} \left[K \|w\|^2 + \sum_{i=1}^{d} (y_i - w^T x_i)^2 \right], \quad K = \frac{\sigma^2}{\tau^2}$$

1c) consider the double derivative along any axis $\Longrightarrow$ Along any axis the function is convex

$$\frac{\partial^2 E}{\partial^2 w_i} = \frac{\partial}{\partial w_i} \left(2\lambda w_i + 2(y - wx)(-x_i) \right)$$

$$= \left(2\lambda + 2x_i^2 \right) > 0$$

COL341: Minor 1	Student Name: CHINMAY MITTAL	Entry Number: 2020CS10336

2. We discussed Naive Bayes classifier in the class. In this question we will discuss Gaussian Naive Bayes, which considers the following data generation model:

   - y is Boolean, and governed by a Bernoulli distribution, with parameter $\theta = \mathbb{P}(y=1)$
   - $\mathbf{x} = \langle x_1, \ldots, x_d \rangle$, where each x_i is a continuous random variable.
   - For each x_i, $\mathbb{P}(x_i \mid y = y_k)$ is a Gaussian distribution, i.e. $\mathbb{P}(x_i \mid y = y_k) = \mathcal{N}(\mu_{ik}, \sigma_i^2)$. Note here we are assuming the standard deviations $\sigma_i i$ vary from feature to feature, but do not depend on y
   - For all i and $j \neq i$, x_i and x_j are conditionally independent given y.

   (a) [5 marks] Show that $\log\left[\frac{\mathbb{P}(x_i|y=0)}{\mathbb{P}(x_i|y=1)}\right] = \frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2} x_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}$

   (b) [10 marks] Show that $\mathbb{P}(y=1 \mid \mathbf{x}) = \frac{1}{1+\exp\left(\log\frac{1-\theta}{\theta} + \sum_{i=1}^{d} \log\left[\frac{\mathbb{P}(x_i|y=0)}{\mathbb{P}(x_i|y=1)}\right]\right)}$

   (c) [5 marks] Recall that in the logistic regression we assume that $\mathbb{P}(y=1 \mid \mathbf{x}) = \frac{1}{1+\exp\left(w_0 + \sum_{i=1}^{d} w_i x_i\right)}$. Using results from part (a) and (b) above, it is easy to see that $\mathbb{P}(y=1 \mid \mathbf{x})$ can be parameterized into the form used by logistic regression under Gaussian Naive Bayes assumption. The above expression gives an alternate way of estimating the value of the weights w_i by estimating the probabilities directly. With the above insight, give advantages and disadvantages of using generative and discriminative estimation of the parameters in logistic regression.

$$N(u_{ik}, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-u_{ik})^2}{2\sigma_i^2}}$$

$$P[x_i|y=0] = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-v_{i0})^2}{2\sigma_i^2}}$$

$$P[x_i|y=1] = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-u_{i1})^2}{2\sigma_i^2}}$$

$$\log\left[\frac{P(x_i|y=0)}{P(x_i|y=1)}\right] = \log_e\left(e^{\frac{(x-v_{i1})^2}{2\sigma_i^2} - \frac{(x-v_{i0})^2}{2\sigma_i^2}}\right)$$

$$= \frac{(x-v_{i1})^2 - (x-v_{i0})^2}{2\sigma_i^2} = \frac{(x^2+v_{i1}^2-2xv_{i1}) - (x^2+v_{i0}^2-2xv_{i0})}{2\sigma_i^2}$$

$$= \frac{v_{i1}^2 - v_{i0}^2}{2\sigma_i^2} + \frac{\cancel{2}x_i(v_{i0}-v_{i1})}{\cancel{2}\sigma_i^2}$$

Q2b) $$p(y=1|x) = \frac{p(y=1)\,p(x|y=1)}{p(y=0)\,p(x|y=0) + p(y=1)\,p(x|y=1)}$$ Bayes rule

$$= \frac{1}{1 + \frac{p(y=0)\prod_{i=1}^{d} p(x_i|y=0)}{p(y=1)\prod_{i=1}^{d} p(x_i|y=1)}}$$ conditional independence

$$= \frac{1}{1 + \exp \log\left(\frac{p(y=0)\prod_{i=1}^{d} p(x_i|y=0)}{p(y=1)\prod_{i=1}^{d} p(x_i|y=1)}\right)}$$

$$= \frac{1}{1 + \exp\left(\log\left(\frac{p(y=0)}{p(y=1)}\right) + \sum_{i=1}^{d}\log\left(\frac{p(x_i|y=0)}{p(x_i|y=1)}\right)\right)}$$

$$= \frac{1}{1 + \exp\left(\log\left(\frac{1-\theta}{\theta}\right) + \sum_{i=1}^{d}\log\left(\frac{p(x_i|y=0)}{p(x_i|y=1)}\right)\right)}$$

Q2c). Discriminative Modelling is better for classification with a lot of datapoints. Since find w_i directly is a much easier and less time consuming process then modelling it indirectly using probabilities.

⇒ With large number of datapoints. discriminative models are preffered (low bias, high variance) → general $w_i x_i + w_0$ instead of a specific form - as in Naive Bayes

With lower number of points they tend to overfit and Generative modelling works much better. because of priors are insert into the model (high bias, low variance) We also get $p(x|y)$ for generating data which is not possible through discriminative modelling.

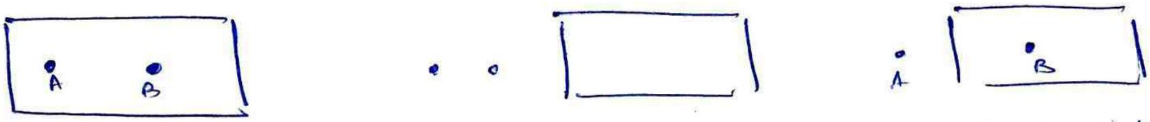
COL341: Minor 1	Student Name: CHINMAY MITTAL	Entry Number: 2020CS10336

3. [10 marks] Let $\mathcal{H}$ consists of all hypotheses in two dimensions $h : \mathbb{R}^2 \to \{+1, -1\}$ that are positive inside some convex set and negative elsewhere (a set is convex if the line segment connecting any two points in the set lies entirely within the set). Compute the VC dimension of $\mathcal{H}$.

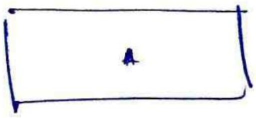
for N = 1

$y = +1$ $\quad$ $y = -1$ $\quad$ shattered

for N = 2

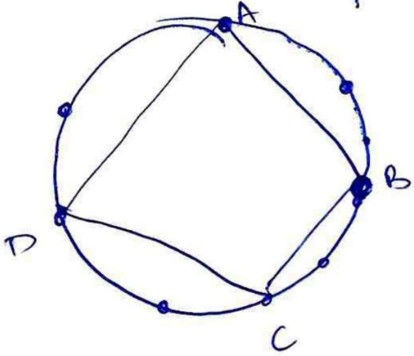


$y = \{+1, +1\}$ $\quad$ $y = \{-1, +1\}$ $\quad$ $y = \{-1, +1\}$



B $\quad y = \{-1, +1\}$ $\quad$ shattered

for N ≥ 3 consider N points which lie on the diameter of a circle.



We show that we can create all dichotimies using $\mathcal{H}$.

in the dichotomy where all points are -ve considerd the empty convex set.

In the dichotomy were 1 point is +ve consider the convex set to be the point itself

When 2 points are +ve consider the convex set to be the line joining the two points.

for ≥ 3 points positive consider The convex polygon joining those points., All other points will become -ve.

This shows that for any N all dichotimies can be generated $\Rightarrow d_{VC} = 2^N$

4. [15 marks] In the class we discussed Perceptron Learning Algorithm (PLA) as follows:

Algorithm 1 Perceptron Learning Algorithm

$w(1) \leftarrow 0$, $t \leftarrow 1$
while any misclassified example left **do**
  Denote the current weight vector as $w(t)$
  Pick any misclassified sample (x_*, y_*) : $\text{sign}(w(t)^\top x_*) \neq y_*$
  $w(t+1) \leftarrow w(t) + y_* x_*$
  $t \leftarrow (t+1)$
end while

Prove that, if the data is linearly separable, then PLA will be able to find one such separator.

Since the data is linearly seperable consider that some optimal w^* exists which clearly classifies all points.

i.e $\text{sign}(w^{*T} x) = y$ for all examples.

consider the angle b/w $w(t)$ and w^* $\longrightarrow \theta(t)$

$$\cos(\theta(t)) = \frac{(w^*)^T w(t)}{\|w^*\| \, \|w(t)\|}$$

Consider the numerator at iteration $t+1$

$$(w^*)^T \left[w(t+1) \right] = (w^*)^T \left[w(t) + y_* x_* \right]$$

$$= (w^*)^T w(t) + \underbrace{y_* (w^*)^T x^*}_{>0}$$

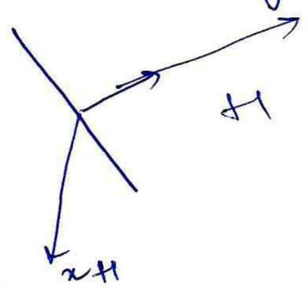
$\Rightarrow$ Numerator ~~increases~~ increases with every iteration.

Since $\text{sign}(w^{*T} x^*) = y_*$

Consider the denominator,

$$\|w^*\| \, \|w(t+1)\| = \|w^*\| \left(\| (w(t) + y^* x^*) \| \right)$$

The angle b/w $w(t)$ & $y^* x^*$ is obtuse since $y^* x^*$ is misclassified by $w(t)$



$\Rightarrow \|w(t) + y^* x^*\| < \|w(t)\|$

The denominator decreases with every iteration.

$\Rightarrow$

$\cos(\theta(t))$ increases with every iteration.

Since $\cos(\theta(t))$ is bounded by 1 it cannot keep increasing in every iteration. $\Rightarrow$ The Algorithm must terminate.

$\Rightarrow$ All examples are correctly classified by $w(t)$ at that point.

$\Rightarrow$ The Algorithm will find a classifier which seperates the classes.

MAY MITTAL 2020CS10336

$$N(\upsilon_{ik}, \sigma_i^2) = A \cdot e^{-\frac{(x-\upsilon_{i0})^2}{2\sigma_1^2}} \qquad \frac{1}{\sqrt{2\pi\sigma^2}}$$

$$A\, e^{-\frac{(x-\upsilon_{i1})^2}{2\sigma^2}}$$

(ROUGH WORK)

$$p(y=1|x) = \frac{p(x|y=1)\,p(y)}{p(x)}$$

$$= \frac{p(x|y=1)\,p(y=1)}{p(x|y=0)\,p(y=1) + p(x|y=1)\,p(y=0)}$$

$$-\log(p(y=1|x)) = \log\left(1 + \frac{p(x|y=0)\,p(y=0)}{p(x|y=1)\,p(y=1)}\right)$$

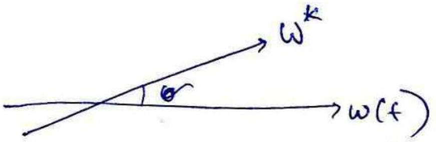
$1-\sigma$

σ

$$p(y=1|x) = \frac{1}{1 + \quad \pi\, p}$$

$p(w)\, p(x|w)$

$-\log p(w)$



$$\cos\theta = \frac{w_*^T\left(w(t) + y_* x_*\right)}{|w^*|\ |w(t)|}$$

$$w_*^T w(t) + \underbrace{w_*^T y_* x_*}_{+ve.}$$

$$|w(t+1)|$$

$$2(y - wx)(-x)$$