DS5220 - Homework 02

KAVANA VENKATESH

1. Linear Regression

Given $\hat{\theta} = \underset{\theta}{\arg\min} \sum_{i=1}^{N} \left(\theta^T \phi(x_i) - y_i\right)^2 + \lambda \|\theta - \alpha\|_2^2$

Using the definition of L_2 norm, we can write the above equation as,

$$\underset{\theta}{\arg\min} \sum_{i=1}^{N} \left(\theta^T \phi(x_i) - y_i\right)^2 + \lambda \|(\theta - \alpha)\|_2^2$$

$$= \| \phi \hat{\theta} - Y\|_2^2 + \lambda \| \hat{\theta} - \alpha\|_2^2$$

where $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$ and $\phi = \begin{bmatrix} \phi(x_1)^T \\ \vdots \\ \phi(x_N)^T \end{bmatrix}$

In order to find the closed form solution of the above equation, take derivative wrt θ and equate it to 0.

$$2\phi^T(\phi\hat{\theta} - Y) + 2\lambda(\hat{\theta} - \alpha) = 0.$$

$$\phi^T(\phi\hat{\theta} - Y) = -\lambda(\hat{\theta} - \alpha)$$

$$\phi^T\phi\hat{\theta} - \phi^T Y = -\lambda\hat{\theta} + \lambda\alpha$$

$$(\phi^T\phi + \lambda I)\hat{\theta} = (\phi^T Y + \lambda\alpha)$$

$$\Rightarrow \boxed{\hat{\theta} = (\phi^T \phi + \lambda I)^{-1} (\phi^T Y + \lambda a)}$$

2. Given the Robust regression model

$$\min_{\theta} \sum_{i=1}^{N} l_\delta (y_i - \theta^T \phi(x_i))$$

2. a) Steps of Batch gradient descent in order to obtain the solution for θ.

$$\text{cost function} = \min_{\theta} \sum_{i=1}^{N} \begin{cases} \frac{1}{2} (y_i - \theta^T \phi(x_i))^2 \\ \delta |y_i - \theta^T \phi(x)| - \frac{1}{2} \delta^2 \end{cases}$$

$$= \min_{\theta} \sum_{i=1}^{N} \begin{cases} \frac{1}{2} (y_i - \phi^T(x_i)\theta)^2 & \text{if } |y_i - \theta^T \phi(x_i)| \le \delta \\ \delta |y_i - \phi(x_i)\theta| - \frac{\delta^2}{2}, & |y_i - \theta^T \phi(x_i)| > \delta \end{cases}$$

Gradient of Huber loss will be,

$$\frac{\partial \text{cost}}{\partial \theta} = \frac{1}{N} \sum_{i=1}^{N} \begin{cases} -\phi(x_i)(y_i - \phi^T(x_i)\theta), & \text{if } |y_i - \theta^T \phi(x_i)| \le \delta \\ -\delta \,\text{sign}(y_i - \phi^T(x_i))\phi(x_i), & \text{if } |y_i - \theta^T \phi(x_i)| > \delta \end{cases}$$

Sign will be +ve where $e \ge 0$.
-ve where $e < 0$.

→ Batch gradient descent

Step 1:- choose a start value θ.

Step 2:- we pick a value for learning rate ρ.

Step 3:- Let's iterate for $t = 1, 2$ and update θ until convergence point.

$$\theta^{t+1} = \theta^{t} - \rho \nabla_{\theta} l_{\delta}(\theta).$$

By substituting this in the previous eq,

$$\theta^{t+1} = \theta^{t} - \rho \left(-\frac{1}{N} \sum_{i=1}^{N} \phi(x_i)(y_i - \phi^T(x_i)\theta)\right); \quad |y_i - \phi^T(x_i)\theta| \leq \delta.$$

$$\theta^{t+1} = \theta^{t} - \rho \left(-\frac{1}{N} \sum_{i=1}^{N} (+ve)(\phi(x_i))\right); \quad y_i - \phi^T(x_i)\theta > \delta.$$

$$\theta^{t+1} = \theta^{t} - \rho \left(-\frac{1}{N} \sum_{i=1}^{N} (-ve)(\phi(x_i))\right); \quad y_i - \phi^T(x_i)\theta < \delta$$

Step 4:- Repeat step 3 until the stopping criteria is met, which is change is norm of the gradient $\leq$ threshold value.

2.B) Stochastic gradient descent

Step 1:- choose the value of θ such that $\theta \in \mathbb{R}^{d+1}$.

Step 2:- pick a value for learning rate ρ.

Step 3: update θ by iterating for $t=1,2$ .... until convergence point

$$\theta^{t+1} = \theta^{t} - \rho \nabla_{\theta} l_{\delta}(x_i, \theta)$$

For random minibatch of size 1, the gradient descent changes as follows.

$$\nabla_{\theta} l_{\delta}(e) = \frac{d}{d\theta} l_{\delta}(e) = \begin{cases} -\frac{1}{2} \cdot 2\phi(x_i)(y_i - \phi^T(x_i)\theta) ; & |y_i - \theta^T\phi(x_i)| \leq \delta \\ -\delta \cdot \text{sign}(y_i - \phi^T(x_i)\theta)\phi(x_i) ; & |y_i - \theta^T\phi(x_i)| > \delta \end{cases}$$

Step 4: we repeat step 3 until convergence is met.

Change in norm of the gradient $\leq$ threshold value

3. Given $P(A|B) + P(A|B^c) = 1 \ \forall \ A, B \subseteq \Omega$

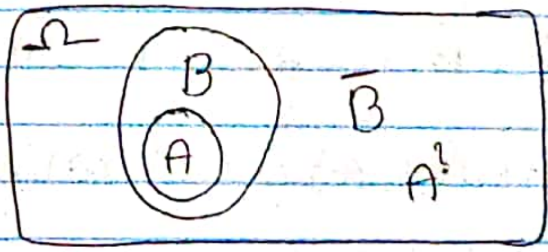
and $0 < P(A) < 1$

3.1) FALSE.

$P(A)$ can be written as $(A|\Omega) = P_\Omega(A)$. This implies that by writing $P(A|B)$ we are looking for possibilities of an event A out of all the outcomes of B.

$$\Rightarrow P(A|B) = P_B(A).$$

Once we look at P_B, we can not move from P_B to $P_{\bar{B}}$. i.e, probabilities in one sample space P_B can not tell us anything about the probabilities in another sample space $P_{\bar{B}}$.



$\Rightarrow$ To disprove the above statement (given) consider the popular example of rolling a fair 6-sided dice.

$P(\text{Even}) = 1/2$, $P(\text{odd}) = 1/2$.

$P(x=2 \mid Even) = 1/3$ and

$P(x=2 \mid \overline{Even}) = P(x=2 \mid odd)$

Since it's a 2 event experiment and there are only 2 possibilities $\Rightarrow$ complement (even) = odd.

Now, $P(x=2 \mid \overline{Even}) = P(x=2 \mid odd) = 0$

$P(x=2 \mid Even) + P(x=2 \mid \overline{Even}) = 1/3 + 0$

$= 1/3 \neq 1$

Hence, $\boxed{P(A \mid B) + P(A \mid \overline{B}) \neq 1}$

3.2) $P(B^c \cap (A \cup B)) + P(A^c \cup B) = 1$

TRUE

<u>Proof:</u> $P(B^c \cap (A \cup B) + P(A^c \cup B)$

$\Rightarrow P((B^c \cap A) \cup (B^c \cap B)) + P(A^c \cup B)$ Distributive Law

$\Rightarrow P(B^c \cap A) + P(B^c \cap B) + P(A^c \cup B)$

But $P(B^c \cap B) = 0$

$\therefore P(B^c \cap A) + 0 + P(B \cup A^c)$

Scanned with CamScanner

$\Rightarrow P(B^c \cap A) + P(B^c \cap A)^c$

$= \underline{\underline{1}} \quad (\because P(A) + P(A') = 1)$

3.3) Given $P(A_1, A_2 \ldots A_n) = P(A_1)P(A_2|A_1) \ldots$
$* P(A_n | A_1, \ldots, A_{n-1})$.

By definition of the conditional proba-bility,

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

Consider the RHS of the equation,

$$P(A_1) P(A_2|A_1) . P(A_3 | A_1 \cap A_2) \ldots P(A_n | A_1 \cap A_2 \ldots A_{n-1})$$

Now apply the definition of the conditi-onal probability to the above eq.

$$P(A_1) . \frac{P(A_2 \cap A_1)}{P(A_1)} . \frac{P(A_3 \cap A_2 \cap A_1)}{P(A_1 \cap A_2)} \ldots \frac{P(A_1 \cap A_2 \ldots \cap A_n)}{P(A_1 \cap A_2 \cap \ldots \cap A_{n-1})}$$

We can notice that every term in the numerator and denominator gets cancelled except the last term

$P(A_1 \cap A_2 \ldots \cap A_n)$, which is nothing but the RHS.

3-4) Given X and Y are discrete independent random variables, then
$E[XY] = E[X]E[Y]$

For discrete random variables X and Y,

$$E(XY) = \sum_i \sum_j x_i y_j f_{xy}(x_i, y_j)$$

$$= \sum_i \sum_j x_i y_j f_x(x_i) f_y(y_j)$$

Since for independent events,
$f_{xy}(x_i, y_j) = f_x(x_i) f_y(y_j)$

$$= \left(\sum_i x_i f_x(x_i)\right)\left(\sum_j y_j f_y(y_j)\right)$$

$$\boxed{E(XY) = E(X)E(Y)}$$

4. $P_\delta(X_i = x_i) = e^{-(\delta^2 + \delta x_i)}$

As the variables $X_1, X_2 \ldots X_N$ are i.i.d,

a) $\mathcal{L}(\delta) = \prod_{i=1}^{N} P_\delta(X_i = x_i) = \prod_{i=1}^{N} e^{-(\delta^2 + \delta x_i)}$

is the likelihood function.

$$\mathcal{L}(\delta) = e^{-(\delta^2+\delta x_1)} \cdot e^{-(\delta^2+\delta x_2)} \cdots e^{-(\delta^2+\delta x_n)}$$

$$= e^{-((\delta^2+\delta^2+ \quad +\delta^2)+\delta(x_1+x_2+ \quad +x_n))}$$

$$\boxed{\mathcal{L}(\delta) = e^{-(N\delta^2+\delta\sum_{i=1}^{N} x_i)}}$$

b). To find the log likelihood function, let's take ln.

$$\ln(\mathcal{L}(\delta)) = \ln(e^{-(N\delta^2+\delta\sum_{i=1}^{N} x_i)})$$

$$= -N\delta^2 - \delta\sum_{i=1}^{N} x_i$$

Taking derivatives wrt δ,

$$\frac{\delta}{\partial\delta}\ln(L(\delta)) = \frac{\partial}{\partial\delta}\left(-N\delta^2-\delta\sum_{i=1}^{N} x_i\right).$$

$$= -2N\delta - \sum_{i=1}^{N} x_i$$

To find the maximum likelihood, let's equate the above eq to 0.

$$\frac{\partial}{\partial\delta}\ln(\mathcal{L}(\delta)) = -2N\delta - \sum_{i=1}^{N} x_i = 0.$$

$$\Rightarrow -2N\delta = +\sum_{i=1}^{N} x_i$$

$$\boxed{\delta^* = \frac{-\sum_{i=1} x_i}{2N}}$$

5) Given $p(y=1|x) = \sigma(w^T x)$

5.1). $w^T x_n < 0.3$.

In order to get the class of the function, let us use the sigmoid function.

$$\sigma(w^T x_n) = \frac{1}{1+e^{-w^T x_n}}$$

Let us now substitute the boundary of $w^T x$ i.e, 0.3

$$\sigma(0.3) = \frac{1}{1+e^{-0.3}} = \frac{1}{1+0.74} \simeq 0.57$$

WKT, for a sigmoid function.

$$y = \begin{cases} 1 & ; \quad p(y=1|x) > 0.5 \\ 0 & ; \quad \text{Otherwise} \end{cases}$$

For $\sigma(z)$ to be equal to exactly $1/2$, $w^T x = 0$.

$$\Rightarrow \boxed{\begin{array}{l} x_n \in \text{class } 1 \text{ if } 0 \le w^T x < 0.3 \\ x_n \in \text{class } 0 \text{ if } w^T x < 0 \end{array}}$$

5.2) Given $\frac{1}{1+e^{\omega^T x}} = 0.7$

Since $\omega^T x$ is +ve, the function is written wrt class 0.

$\therefore \frac{1}{1+e^{\omega^T x}} = 0.7$

$1+e^{\omega^T x} = \frac{1}{0.7} = \frac{1}{0.7}$

$e^{\omega^T x} = \frac{1}{0.7} - 1$

$\Rightarrow e^{\omega^T x} = \frac{0.3}{0.7}$

$\Rightarrow e^{-\omega^T x} = 7/3$

$\therefore \frac{1}{1+e^{-\omega^T x}} = \frac{1}{1+7/3} = \frac{3}{10} = 0.3$

Since $\sigma(z) = \frac{1}{1+e^{-\omega^T x}} = 0.3 < 0.5$,

$\Rightarrow x_n \in$ class 0

So, the model belongs to class 0 with 70% probability.

6) MLE estimates θ_j^y $j = 0, 1$ as well

6.1) as $\theta_{\bar{x}_l}^{x_l}\Big|_j^y$, $j = 0, 1$ and $\bar{x}_l = 0, 1$ and for $l = 1, 2$.

a) θ_j^y for $j = 0, 1$

$$P(y=0) = 3/7, \quad P(y=1) = 4/7$$

b)

$P(x_1 = 0 \mid y = 0) = 2/3$	$P(x_1 = 0 \mid y = 1) = 2/4$
$P(x_2 = 0 \mid y = 0) = 2/3$	$P(x_2 = 0 \mid y = 1) = 2/4$
$P(x_1 = 1 \mid y = 0) = 1/3$	$P(x_1 = 1 \mid y = 1) = 2/4$
$P(x_2 = 1 \mid y = 0) = 1/3$	$P(x_2 = 1 \mid y = 1) = 2/4$

6.2) $P(y = 0 \mid x_1 = 0, x_2 = 1)$

$$= P(y=0) \cdot P(x_1 = 0 \mid y = 0) \cdot P(x_2 = 1 \mid y = 0)$$

$$= 2/3 \cdot 1/3 \cdot 3/7 = \underline{2/21}$$

6.3). Solution of $p(y=0 \mid x_1=0, x_2=1)$ without Naive Bayes is.

$$P(x_1=0, x_2=1 \mid y=0) \cdot P(y=0)$$

$$= \frac{1}{3} \cdot \frac{3}{7}$$

$$= \frac{1}{7}$$

Also, without the Naive Bayes assumption, we can directly look at the given table where in $x_1=0$ and $x_2=1$ co-exist and write the $p(y=0 \mid x_1=0, x_2=1)$ as,

$$P(y=0 \mid x_1=0, x_2=1) = \frac{1}{7}$$