Assignment 4
AI2000
Foundations of Machine Learning

Darpan Gaur
CO21BTECH11004

Problem 1

Consider case of shattering 3 points, x_1, x_2, x_3. Say, $x_1 < x_2 < x_3$, and x_1 and x_3 are classified as 1. Give labelling x_2 as 0. But as $x_1 < x_2 < x_3$, we can't classify x_2 as 0. Hence, 3 points can't be shattered, so VC dimension is 2.

Problem 2

Given,

$$y(x, w) = w_0 + \sum_{k=1}^{D} w_k x_k = x^T w$$

Now adding gaussian noise $\epsilon_k \sim \mathcal{N}(0, \sigma^2)$ to x_k:

$$y(x, w) = w_0 + \sum_{k=1}^{D} w_k (x_k + \epsilon_k) = (x + \epsilon)^T w$$

Sum of squared loss:

$$E(w) = \frac{1}{2} \sum_{i=1}^{N} (y(x_i, w) - t_i)^2$$

Analyzing the effect of noise on $E(w)$, we see expected squared loss, where expectaion taken over ϵ:

$$L(w) = \mathbb{E}_\epsilon [\frac{1}{2} \sum_{i=1}^{N} ((x_i + \epsilon_i)^T w - t_i)^2] = \frac{1}{2} \sum_{i=1}^{N} \mathbb{E}_\epsilon [((x_i^T w - t_i) + \epsilon_i^T w)^2]$$

$$\implies L(w) = \frac{1}{2} \sum_{i=1}^{N} \mathbb{E}_\epsilon [(x_i^T w - t_i)^2 + 2\epsilon_i^T w (x_i^T w - t_i) + w^T \epsilon_i \epsilon_i^T w]$$

$$\implies L(w) = \frac{1}{2}\sum_{i=1}^{N}[(x_i^T w - t_i)^2 + 2\,\mathbb{E}_\epsilon(\epsilon_i^T) w (x_i^T w - t_i) + w^T\,\mathbb{E}_\epsilon(\epsilon_i \epsilon_i^T) w]$$

As $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, we have $\mathbb{E}_\epsilon(\epsilon_i) = 0$ and $\mathbb{E}_\epsilon(\epsilon_i \epsilon_i^T) = \sigma^2 I$.

$$\implies L(w) = \frac{1}{2}\sum_{i=1}^{N}[(x_i^T w - t_i)^2 + w^T \sigma^2 I w]$$

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}(x_i^T w - t_i)^2 + \frac{\sigma^2}{2} w^T w \tag{1}$$

For $L2$ regularization,

$$L(w) = \frac{1}{2}\sum_{i=1}^{N}(x_i^T w - t_i)^2 + \lambda w^T w \tag{2}$$

By comparing (1) and (2), we see that adding gaussian noise to input is equivalent to $L2$ regularization with $\lambda = \frac{\sigma^2}{2}$.

Problem 3

Part (a)

Dendogram for the final result of hierarchical clustering with single linkage is shown in Figure 1.

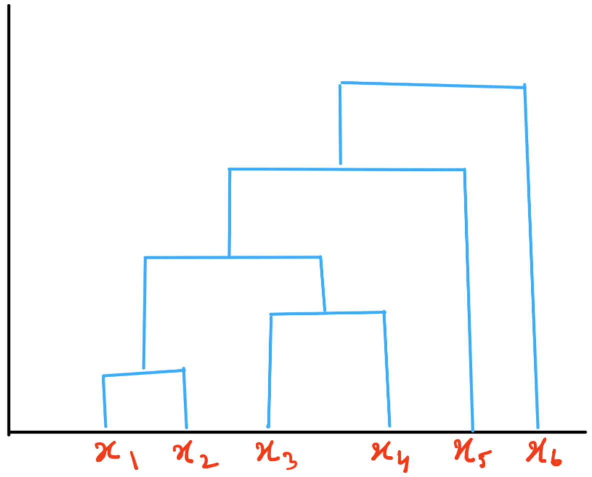

Figure 1: Dendogram for hierarchical clustering with single linkage

Part (b)

Dendogram for the final result of hierarchical clustering with complete linkage is shown in Figure 2.

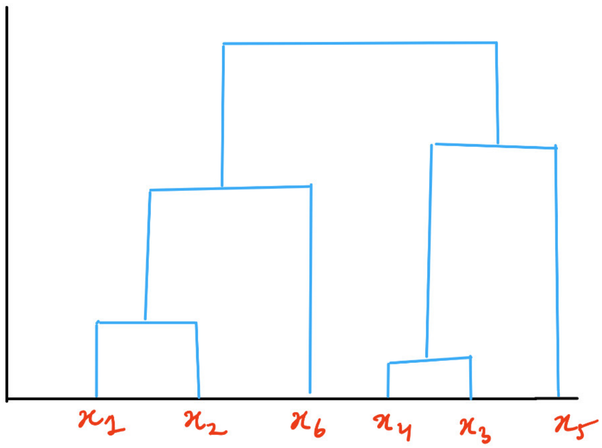

Figure 2: Dendogram for hierarchical clustering with complete linkage

Part (c)

Change $dist(x_1, x_2)$ to 0.13 and $dist(x_3, x_6)$ to 0.95.
This will not change the resluts of the above question as samll pertubation given to smallest and largest distance, to decrease and increase respectively will not change the results of hierarchical clustering.

Problem 4

Part (a)

$$C = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$$

$$\implies C = \mathbb{E}[XX^T - X\,\mathbb{E}[X]^T - \mathbb{E}[X]X^T + \mathbb{E}[X]\,\mathbb{E}[X]^T]$$

As $x = a\delta_k = a \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$, where a is scalar and δ_k is k-th unit vector, distributed over $1, \dots, M$.

$$\implies C_{ij} = \mathbb{E}[a^2 \delta_i \delta_j^T] - a^2 \delta_i\,\mathbb{E}[\delta_j^T] - a^2\,\mathbb{E}[\delta_i]\delta_j^T + a^2\,\mathbb{E}[\delta_i \delta_j^T]$$

Two cases arise:

- $i = j$:

$$C_{ii} = \mathbb{E}[a^2 \delta_i \delta_i^T] - a^2\,\mathbb{E}[\delta_i]\,\mathbb{E}[\delta_i^T]$$

$$\implies C_{ii} = \frac{\mathbb{E}[a^2]}{M} - \frac{E[a]^2}{M^2}$$

- $i \neq j$:

$$C_{ij} = -\,\mathbb{E}[a\delta_i]\,\mathbb{E}[a\delta_j^T]$$

$$\implies C_{ij} = -\frac{E[a]^2}{M^2}$$

Combinining both cases, we get:

$$C = \frac{\mathbb{E}[a^2]}{M} - \frac{E[a]^2}{M^2}$$

Gives, $\lambda = -\frac{E[a]^2}{M^2}, \mu = \frac{\mathbb{E}[a^2]}{M}$.

Part (c)

PCA is not a good way to select features for this problem.
This is because, covariance matrix would have one large eigenvalue and $M - 1$ small eigenvalues. So, variance along the direction of large eigenvalue is large, hence PCA would select that. But, it may not be the best feature for classification.

Problem 5

Part (a)

Refer to the code in `5_logisticRegression.ipynb`.

Part (b)

i

Logistic model $P(\mathrm{y} = 1|x_1, x_2)$ is given by:

$$P(\mathrm{y} = 1|x_1, x_2) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}}$$

Cross entropy loss is given by:

$$CE(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log P(\mathrm{y} = 1|x_1, x_2) + (1 - y^{(i)}) \log\left(1 - P(\mathrm{y} = 1|x_1, x_2)\right)$$

ii

The updated paramters for logistic regression after one iteration are:

$$\theta_0 = 1.00316, \theta_1 = 1.50535, \theta_2 = 0.50196$$

iii

At the convergence of gradient descent, metrics are:

- Accuracy: 0.67
- Precision: 0.6
- Recall: 1.0

Problem 6

Refer to the code in `6_house_prices.ipynb`.
Model used and scores on house price prediction are:

Model	Score
CatBoost	**0.13617**
LightGBM	**0.14359**
Random Forest	0.14851
XGBoost	0.15374
Linear Regression	0.35652
SVM Regression	0.41645

Top 2 models are CatBoost and LightGBM.

- CatBoost: Gradient boosting model, which uses symmetric trees. It is robust to overfitting and can handle categorical features well by using ordered encoding.
- LightGBM: Gradient boosting model, which uses leaf-wise tree growth. Uses gradient-based one-side sampling, exclusive feature bundling and histogram-based algorithms.
- Random forest and XGBoost also give similar performance as compared to CatBoost and LightGBM. As both similare to them in using ensemble of trees.
- Linear regression and SVM regression give poor performance as they are linear models and can't capture non-linear relationships in data.