Machine Learning
Home Assignment 3

Carlo Rosso rkm957

Contents

1 Preprocessing (33 points)

1.1 Importance of Preprocessing (6 points)

We have the following data points:

1.1.1 a)

Person	Age in years	Income in thousands of USD	Paied off
A	47	35	yes
B	22	40	no
C	21	36	-

$$d_A = \sum_{i=1}^{2} (x_i - y_i)^2 = (21 - 47)^2 + (36 - 35)^2 = 677 \tag{1}$$

$$d_B = (21 - 22)^2 + (36 - 40)^2 = 17 \tag{2}$$

Therefore we get $d_A > d_B$ and we can conclude the BoL should not give credit to C, according to the nearest neighbor algorithm.

1.1.2 b)

Person	Age in years	Income in USD	Paied off
A	47	35000	yes
B	22	40000	no
C	21	36000	-

$$d_A = \sum_{i=1}^{2} (x_i - y_i)^2 = (21 - 47)^2 + (36000 - 35000)^2 = 1000676 \tag{3}$$

$$d_B = (21 - 22)^2 + (36000 - 40000)^2 = 16000001 \tag{4}$$

Therefore we get $d_A < d_B$ and we can conclude the BoL should give credit to C, according to the nearest neighbor algorithm.

1.2 Input Centering (9 points)

1.2.1 a)

Considering the following equations:

$$z_n = x_n - \bar{x}, \forall n = 1, ..., N \tag{5}$$

$$\bar{x} = \frac{1}{N} X^T 1 \tag{6}$$

$$\gamma = 1 - \frac{1}{N} 11^T \tag{7}$$

We can show that:

$$Z = \gamma X \tag{8}$$

Indeed:

$$z_n = x_n - \bar{x} \forall n = 1, ..., N \tag{9}$$

Turning this into a matrix form, we get:

$$\begin{aligned} Z &= X - 1\bar{x}^T \\ &= X - 1\left(\frac{1}{N} X^T 1\right)^T \end{aligned} \tag{10}$$

Remembering that $(AB)^T = B^T A^T$, we get:

$$\begin{aligned} Z &= X - \frac{1}{N} 11^T X \\ &= IX - \frac{1}{N} 11^T X \\ &= \left(I - \frac{1}{N} 11^T\right) X \\ &= \gamma X \end{aligned} \tag{11}$$

Therefore, we have shown that $Z = \gamma X$.

1.2.2 b)

Considering that Z is a $N \times D$ matrix and $\text{rank}(Z) = \text{rank}(Z^T)$.

Citing the rank-nullity theorem[1], we have that given a matrix of dimension $d \times d$ A:

$$\text{rank}(A) + \text{rank}(\ker(A)) = d. \tag{12}$$

Citing a property of the rank[2], given two matrixes A, B:

$$\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B)). \tag{13}$$

Therefore, we have that $\text{rank}(\gamma) + \text{rank}(\ker(\gamma)) = N$. And since $\text{rank}(\ker(\gamma)) = 1$[3], we have $\text{rank}(\gamma) = N - 1$.

Therefore, $\text{rank}(Z) = \text{rank}(\gamma X) \leq \min(\text{rank}(X), N - 1) < N$.

1.3 Input Whitening (18 points)

1.3.1 a)

Given the following:

$$\text{Var}(\hat{x}_1) = \text{Var}(\hat{x}_2) = 1 \tag{14}$$

$$\mathbb{E}[\hat{x}_1] = \mathbb{E}[\hat{x}_2] = 0 \tag{15}$$

$$x_1 = \hat{x}_1 \tag{16}$$

$$x_2 = \sqrt{1 - \varepsilon^2}\hat{x}_1 + \varepsilon\hat{x}_2 \text{ for } \varepsilon \in [-1, 1] \tag{17}$$

$$\text{Cov}(\hat{x}_1, \hat{x}_2) = 0 \tag{18}$$

The last equation is given by the fact that the two variables are independent.

Therefore we already have the variance of x_1: $\text{Var}(x_1) = \text{Var}(\hat{x}_1) = 1$.

The variance of x_2 is given by:

$$\begin{aligned} \text{Var}(x_2) &= \sqrt{1 - \varepsilon^2}^2 \ \text{Var}(\hat{x}_1) + \varepsilon^2 \ \text{Var}(\hat{x}_2) \\ &= 1 - \varepsilon^2 + \varepsilon^2 \\ &= 1 \end{aligned} \tag{19}$$

Finally, the covariance between x_1 and x_2 is given by:

$$\begin{aligned} \text{Cov}(x_1, x_2) &= \sqrt{1 - \varepsilon^2} \ \text{Cov}(\hat{x}_1, \hat{x}_1) + \varepsilon \ \text{Cov}(\hat{x}_1, \hat{x}_2) \\ &= \sqrt{1 - \varepsilon^2} \ \text{Var}(\hat{x}_1) \\ &= \sqrt{1 - \varepsilon^2} \end{aligned} \tag{20}$$

1.3.2 b)

Given the following:

$$x = (x_1, x_2)^T \tag{21}$$

$$\hat{x} = (\hat{x}_1, \hat{x}_2)^T \tag{22}$$

$$f(\hat{x}) = \hat{w}_1\hat{x}_1 + \hat{w}_2\hat{x}_2 \tag{23}$$

Follows equivalent statements one below the other:

$$w_1x_1 + w_2x_2 = \hat{w}_1\hat{x}_1 + \hat{w}_2\hat{x}_2 \tag{24}$$

$$w_1\hat{x}_1 + w_2\left(\sqrt{1-\varepsilon^2}\hat{x}_1 + \varepsilon\hat{x}_2\right) = \hat{w}_1\hat{x}_1 + \hat{w}_2\hat{x}_2 \tag{25}$$

$$w_1\hat{x}_1 + w_2\sqrt{1-\varepsilon^2}\hat{x}_1 + w_2\varepsilon\hat{x}_2 = \hat{w}_1\hat{x}_1 + \hat{w}_2\hat{x}_2 \tag{26}$$

$$\begin{cases} \left(w_1 + w_2\sqrt{1-\varepsilon^2}\right)\hat{x}_1 = \hat{w}_1\hat{x}_1 \\ w_2\varepsilon\hat{x}_2 = \hat{w}_2\hat{x}_2 \end{cases} \tag{27}$$

$$\begin{cases} w_1 + w_2\sqrt{1-\varepsilon^2} = \hat{w}_1 \\ w_2\varepsilon = \hat{w}_2 \end{cases} \tag{28}$$

And so we we arrive to the final conclusion that f is linear in the correlated inputs:

$$\begin{cases} w_1 = \hat{w}_1 - \frac{\hat{w}_2}{\varepsilon}\sqrt{1-\varepsilon^2} \\ w_2 = \frac{\hat{w}_2}{\varepsilon} \end{cases} \tag{29}$$

1.3.3 c)

Given target function:

$$f(\hat{x}) = \hat{x}_1 + \hat{x}_2 \tag{30}$$

The constraint C:

$$w_1^2 + w_2^2 \leq C \tag{31}$$

If we perform regression with the correlated inputs x, then let's find the minimum value of C such that the constraint is satisfied. First of all let's compute the values of w_1 and w_2 considering the previous results:

$$\begin{aligned} \hat{w}_1 &= 1 \\ \hat{w}_2 &= 1 \end{aligned} \tag{32}$$

Therefore:

$$
\begin{aligned}
w_1 &= 1 - \frac{1}{\varepsilon}\sqrt{1-\varepsilon^2} \\
w_2 &= \frac{1}{\varepsilon}
\end{aligned}
\tag{33}
$$

Now we can compute the value of C:

$$
\begin{aligned}
C &= w_1^2 + w_2^2 \\
&= \left(1 - \frac{1}{\varepsilon}\sqrt{1-\varepsilon^2}\right)^2 + \left(\frac{1}{\varepsilon}\right)^2 \\
&= 1 + \frac{1-\varepsilon^2}{\varepsilon^2} - \frac{2}{\varepsilon}\sqrt{1-\varepsilon^2} + \frac{1}{\varepsilon^2} \\
&= \frac{2}{\varepsilon^2} - \frac{2\sqrt{1-\varepsilon^2}}{\varepsilon}
\end{aligned}
\tag{34}
$$

1.3.4 d)

Let's compute the following limit:

$$
\lim_{\varepsilon \to 0} C = \lim_{\varepsilon \to 0}\left(\frac{2}{\varepsilon^2} - \frac{2\sqrt{1-\varepsilon^2}}{\varepsilon}\right) = \infty \tag{35}
$$

2 Competition Design to Find Defective Products (24 points)

2.1

Follows the theorem of generalization bound for selection from finite $\mathcal{H}$:

$$
\mathbb{P}\left(L\left(\hat{h}_S^*\right) \leq \hat{L}\left(\hat{h}_S^*, S\right) + \sqrt{\frac{\ln\left(\frac{M}{\delta}\right)}{2n}}\right) \geq 1 - \delta \tag{36}
$$

Let's repeat our hypothesis:

$$
M = 20 \tag{37}
$$

$$
\delta = 2 \tag{38}
$$

We are looking for the minimum value of n such that the follwing inequality is satisfied:

$$\sqrt{\frac{\ln\left(\frac{M}{\delta}\right)}{2n}} \leq 0.04 \tag{39}$$

Therefore:

$$\frac{\ln\left(\frac{20}{2}\right)}{2 \cdot 0.04^2} = 312.5 < 313 = n \tag{40}$$

2.2

Given the following:

$$n = 1800 \tag{41}$$

$$\delta = 2 \tag{42}$$

We are looking for the maximum value of M such that the following inequality is satisfied:

$$\sqrt{\frac{\ln\left(\frac{M}{\delta}\right)}{2n}} \leq 0.04 \tag{43}$$

Therefore:

$$2\exp(0.04^2 \cdot 2 \cdot 1800) \sim 634.7 > 634 = M \tag{44}$$

3 Combining Multiple Confidence Intervals (22 points)

Given the following:

$$\begin{aligned} i \in I &= \{1, 2, 3\} \\ S_i &= S \\ \mathrm{CI}_i &= [l_i, u_i] \qquad \text{w.p. } 1 - \delta_i \\ 0.99 &= \prod_I (1 - \delta_i) \\ \delta &= \delta_i = \delta_j \qquad \forall i, j \in I \end{aligned} \tag{45}$$

Let's compute the value of δ:

$$1 - \sqrt[3]{0.99} \sim 0.0033 < 0.004 = \delta \tag{46}$$

We could compute a more precise value for δ by I only need to show how do answer this question.

Alex can choose any combination of the confidence intervals endpoints such that

$l_{\text{chosen}} \leq u_{\text{chosen}}$, because any such combination is a valid (at least 99)-CI. Therefore, he should choose the combination that minimizes the length of the CI:

$$\text{CI} = [\max(l_i), \min(u_i)] \tag{47}$$

4 Early Stopping (21 points)

4.1 Neural network with early stopping (21 points)

> Statistical bias, in the mathematical field of statistics, is a systematic tendency in which the methods used to gather data and generate statistics present an inaccurate, skewed or biased depiction of reality.
>
> — Wikipedia[4]

4.1.1 Predefined Stopping

The S_{val} has no influence on the choice of the target function h_{t^*}, so the bias is not present.

4.1.2 Non-adaptive Stopping

It is chosen the target function h_{t^*} that minimizes the validation error $\hat{L}(h_{t^*})$. Therefore, the dataset is used to choose the best target function, which lead the final model to be biased by the validation set S_{val}.

4.1.3 Adaptive Stopping

h_{t^*} is chosen in the sequence of hypotesis $h_1, h_2, h_3, ..., h_t$. While the target function is not chosen based on the validation set, the sequence stops when the validation does not improve anymore for a certain number of steps. Therefore, the sequence of models is biased by the validation set S_{val}.

As a counterexample to the claim that the bias is not present, let's consider the case in which a different validation set S'_{val} is used. Then, it would produce the sequence of hypotesis models $h_1, h_2, h_3, ..., h_j$ and $j \neq t$. Let $j > t$ and h_j be the best model, thus the final model choice differs from the one obtained with the original validation set S_{val}. Therefore, we can conclude that the final model is biased by the validation set S_{val}.

4.2

I have already cited the theorem of generalization bound (see Equation 36), so follows the solution for the two cases.

4.2.1 Predefined Stopping

We have $M = 1$, because we are only considering the final model:

$$\mathbb{P}\left(L\left(\hat{h}_S^*\right) \leq \hat{L}\left(\hat{h}_S^*, S\right) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\right) \geq 1 - \delta \tag{48}$$

4.2.2 Non-adaptive Stopping

We have $M = T$, where T is the number of epochs and so the number of models to consider:

$$\mathbb{P}\left(L\left(\hat{h}_S^*\right) \leq \hat{L}\left(\hat{h}_S^*, S\right) + \sqrt{\frac{\ln\left(\frac{T}{\delta}\right)}{2n}}\right) \geq 1 - \delta \tag{49}$$

Bibliography

[1] Wikipedia contributors, "Rank–nullity theorem — Wikipedia, The Free Encyclopedia." [Online]. Available: https://en.wikipedia.org/w/index.php?title=Rank%E2%80%93nullity_theorem&oldid=1219582126

[2] Wikipedia contributors, "Rank (linear algebra) — Wikipedia, The Free Encyclopedia." [Online]. Available: https://en.wikipedia.org/w/index.php?title=Rank_(linear_algebra)&oldid=1245285666

[3] Wikipedia contributors, "Centering matrix — Wikipedia, The Free Encyclopedia." [Online]. Available: https://en.wikipedia.org/w/index.php?title=Centering_matrix&oldid=1242793579

[4] Wikipedia contributors, "Bias (statistics) — Wikipedia, The Free Encyclopedia." [Online]. Available: https://en.wikipedia.org/w/index.php?title=Bias_(statistics)&oldid=1225782713