Home Assignment 2

Yu Wang(ndp689)

December 7, 2021

Contents

1 Preprocessing

(a) variance(x_1) = 1;

variance(x_2) = D($\sqrt{1-\epsilon^2}\hat{x}_1$) + D($\epsilon\hat{x}_2$)= $(1-\epsilon^2)D(\hat{x}_1) + \epsilon^2 D(\hat{x}_2) = 1$;

We can know that $x_2 = \sqrt{1-\epsilon^2}x_1 + \epsilon\hat{x}_2$. So, x_1 and x_2 are not independent. Then, we have:

$$\begin{aligned} covariance(x_1, x_2) &= \frac{D(x_1 + x_2) - D(x_1) - D(x_2)}{2} \\ &= \frac{(\sqrt{1-\epsilon^2}+1)^2 D(\hat{x}_1) + \epsilon^2 D(\hat{x}_2) - 2}{2} \\ &= \sqrt{1-\epsilon^2} \end{aligned}$$

(b)

$$\begin{aligned} f(x) &= w_1 x_1 + w_2 x_2 \\ &= w_1\hat{x}_1 + w_2(\sqrt{1-\epsilon^2}\hat{x}_1 + \epsilon\hat{x}_2) \\ &= (w_1 + \sqrt{1-\epsilon^2}w_2)\hat{x}_1 + \epsilon w_2\hat{x}_2. \end{aligned}$$

If we let $\hat{w}_1 = w_1 + \sqrt{1-\epsilon^2}w_2, \hat{w}_2 = \epsilon w_2$, we have,

$$\begin{aligned} f(x) &= (w_1 + \sqrt{1-\epsilon^2}w_2)\hat{x}_1 + \epsilon w_2\hat{x}_2 \\ &= \hat{w}_1\hat{x}_1 + \hat{w}_2\hat{x}_2 \end{aligned}$$

So, f is linear in x_1, x_2.

(c) From problem (b), let $\hat{w}_1 = \hat{w}_2 = 1$.
So, $w_1 = \frac{\epsilon-\sqrt{1-\epsilon^2}}{\epsilon}, w_2 = \frac{1}{\epsilon}$. Then,

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

From equation (1), obviously, $\epsilon \in [-1, 0) \cup (0, 1]$. So, let $\epsilon = cos\theta, \theta \in$

$[0, \frac{\pi}{2}) \cup (\frac{\pi}{2}, \pi]$. Then, according equation (1), we have,

$$\begin{aligned} w_1^2 + w_2^2 &= \frac{2}{\epsilon^2} - \frac{2\sqrt{1-\epsilon^2}}{\epsilon} \\ &= 2sec^2\theta - 2tan\theta \\ &= 2(tan\theta - \frac{1}{2})^2 + \frac{3}{2} \\ &\geq \frac{3}{2} \end{aligned}$$

So, minimum value of C is $\frac{3}{2}$

(d) As $\epsilon \to 0$, the minimum $C \to \infty$. So, we need to use a $C \to \infty$ to be able to implement the target function, which is impossible.

2 Logistic regression

2.1 Cross-entropy error measure

(a) We know that we need to maximize the likelihood $\prod_{n=1}^{N} P(y_n|x_n)$. It is equivalent to maximize $\sum_{n=1}^{N} \ln(P(y_n|x_n))$, or minimize $-\sum_{n=1}^{N} \ln(P(y_n|x_n))$. And according to the equation:

$$P(y_n|x_n) = \begin{cases} h(x_n) & y_n = +1 \\ 1 - h(x_n) & y_n = -1 \end{cases}$$

We have,

$$\begin{aligned} E_{in}(w) &= -\sum_{n=1}^{N} \ln(P(y_n|x_n)) \\ &= -\sum_{n=1}^{N} ([\![y_n = +1]\!] \ln(h(x_n)) + [\![y_n = -1]\!] \ln(1 - h(x_n))) \quad (2) \\ &= \sum_{n=1}^{N} \left([\![y_n = +1]\!] \ln \frac{1}{h(x_n)} + [\![y_n = -1]\!] \ln \frac{1}{1 - h(x_n)} \right) \end{aligned}$$

(b) We know that $h(x) = \theta(w^T x)$, so, $\ln \frac{1}{h(x_n)} = \ln(1 + e^{-w^T x_n})$ and $\ln \frac{1}{(1-h(x_n))} = \ln(1 + e^{w^T x_n})$. Then, according to the equation (2), we have,

$$\begin{aligned}
E_{in}(w) &= \sum_{n=1}^{N} \left([\![y_n = +1]\!] \ln \frac{1}{h(x_n)} + [\![y_n = -1]\!] \ln \frac{1}{1-h(x_n)} \right) \\
&= \sum_{n=1}^{N} \left([\![y_n = +1]\!] \ln(1 + e^{-w^T x_n}) + [\![y_n = -1]\!] \ln(1 + e^{w^T x_n}) \right) \\
&= \sum_{n=1}^{N} \ln(1 + e^{-y_n w^T x_n})
\end{aligned}$$

Therefore, minimizing the in-sample error in part (a) is equivalent to minimizing the one in equation (3.9)

2.2 Logistic regression loss gradient

(1) Assuming labels in {-1, 1}

We know that $E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} \ln(1 + e^{-y_n w^T x_n})$, so, we have the gradient,

$$\begin{aligned}
\nabla E_{in}(w) &= -\frac{1}{N} \sum_{n=1}^{N} \frac{y_n x_n e^{-y_n w^T x_n}}{1 + e^{-y_n w^T x_n}} \\
&= -\frac{1}{N} \sum_{n=1}^{N} \frac{y_n x_n}{1 + e^{y_n w^T x_n}} \\
&= \frac{1}{N} \sum_{n=1}^{N} -y_n x_n \theta(-y_n w^T x_n)
\end{aligned}$$

According to the gradient, we know that if an example is misclassified, $y_n w^T x_n$<0, so $\theta(-y_n w^T x_n)$>0.5, while if an example is correctly classified, $y_n w^T x_n$>0, so $\theta(-y_n w^T x_n)$<0.5.

So, the contribution of 'misclassified' example is more to the gradient than a correctly classified one.

(2) Assuming labels in $\{0, 1\}$

We can get the $E_{in}(w)$,

$$\begin{aligned}
\nabla E_{in}(w) &= -\sum_{n=1}^{N} \ln(P(y_n|x_n)) \\
&= -\sum_{n=1}^{N} \left([\![y_n = +1]\!] \ln(h(x_n)) + [\![y_n = 0]\!] \ln(1 - h(x_n))\right) \\
&= -\sum_{n=1}^{N} \left(y_n \ln(h(x_n)) + (1 - y_n) \ln(1 - h(x_n))\right) \\
&= -\sum_{n=1}^{N} \left(y_n(\ln(h(x_n)) - \ln(1 - h(x_n))) + \ln(1 - h(x_n))\right) \\
&= -\sum_{n=1}^{N} \left(y_n \ln \frac{h(x_n)}{1 - h(x_n)} + \ln(1 - h(x_n))\right) \\
&= -\sum_{n=1}^{N} \left(y_n w^T x_n - \ln(1 + e^{w^T x_n})\right)
\end{aligned}$$

And then, its gradient is,

$$\begin{aligned}
\nabla E_{in}(w) &= -\frac{1}{N} \sum_{n=1}^{N} \left(y_n x_n - \frac{x_n e^{w^T x_n}}{1 + e^{w^T x_n}}\right) \\
&= -\frac{1}{N} \sum_{n=1}^{N} (y_n - \theta(w^T x_n)) x_n
\end{aligned}$$

According to the gradient, we know that if an example is misclassified, $|y_n - \theta(w^T x_n)|$ is larger, while if an example is correctly classified, $|y_n - \theta(w^T x_n)|$ is lesser.

So, the contribution of 'misclassified' example is more to the gradient than a correctly classified one.

2.3 Log-odds

Let p = P(Y = 1 | X = x) = $\sigma(w^T x + b)$, so, 1 - p = P(Y = 0 | X = x). Then, we have

$$\begin{aligned}
w^T x + b &= \ln \frac{p}{1 - p} \\
&= \ln \frac{\sigma(w^T x + b)}{1 - \sigma(w^T x + b)}
\end{aligned}$$

This equals:

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

$$\begin{aligned} 1 + e^{w^T x+b} &= 1 + \frac{\sigma(w^T x + b)}{1 - \sigma(w^T x + b)} \\ &= \frac{1}{1 - \sigma(w^T x + b)} \iff \end{aligned}$$

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

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

Then, σ is the logistic function.

2.4 Variable importance

1. How many solutions (i.e., optimal values for the coefficients) would the linear regression optimization problem (without regularization) have if the one-hot encoding was used? Why?

   We will have infinite solutions.

   Reasons:
   If we use the one-hot encoding, the model will be like:

   $$y = w_0 + w_1 gre + w_2 gpa + w_3 rank_1 + w_4 rank_2 + w_5 rank_3 + w_6 rank_4$$

   Where $rank_1 + rank_2 + rank_3 + rank_4 = 1$. So, we cannot avoid the linear dependency and such 4 parameters ($rank_1, ..., rank_4$) are highly relevant. Then, there will be infinite solutions.

2. Why would it be difficult to interpret the variable importance if the one-hot encoding was used?

```
Optimization terminated successfully.
         Current function value: 0.573147
         Iterations 11
                          Results: Logit
=======================================================================
Model:               Logit              Pseudo R-squared:   0.083
Dependent Variable:  admit              AIC:                470.5175
Date:                2021-12-07 15:48   BIC:                494.4663
No. Observations:    400                Log-Likelihood:     -229.26
Df Model:            5                  LL-Null:            -249.99
Df Residuals:        394                LLR p-value:        7.5782e-08
Converged:           1.0000             Scale:              1.0000
No. Iterations:      11.0000
-----------------------------------------------------------------------
        Coef.    Std.Err.       z    P>|z|      [0.025        0.975]
-----------------------------------------------------------------------
const  -3.9054 8947848.5333 -0.0000 1.0000 -17537464.7699 17537456.9590
gre     0.0023       0.0011  2.0699 0.0385         0.0001        0.0044
gpa     0.8040       0.3318  2.4231 0.0154         0.1537        1.4544
rank_1 -0.0846 8947848.5333 -0.0000 1.0000 -17537460.9490 17537460.7799
rank_2 -0.7600 8947848.5333 -0.0000 1.0000 -17537461.6245 17537460.1044
rank_3 -1.4248 8947848.5333 -0.0000 1.0000 -17537462.2892 17537459.4397
rank_4 -1.6360 8947848.5333 -0.0000 1.0000 -17537462.5005 17537459.2284
=======================================================================
```

Figure 1: the model of using the one-hot encoding

From Figure 1, if we use the one-hot encoding, the coefficients of $rank_1, rank_2$, $rank_3$ and $rank_4$ may not be true (because their Std.Err. are very large), and from the %5 significance level, they are not statistically significant.

Why C1 variables are used?

Using C-1 variables is enough, because we can use (1, 0, 0), (0, 1, 0) and (0, 0, 1) to express $rank_2, rank_3, rank_4$ respectively, and then, we can also use (0, 0, 0) to express $rank_1$.

```
Optimization terminated successfully.
         Current function value: 0.573147
         Iterations 6
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.083
Dependent Variable: admit            AIC:              470.5175
Date:               2021-12-07 16:54 BIC:              494.4663
No. Observations:   400              Log-Likelihood:   -229.26
Df Model:           5                LL-Null:          -249.99
Df Residuals:       394              LLR p-value:      7.5782e-08
Converged:          1.0000           Scale:            1.0000
No. Iterations:     6.0000
-------------------------------------------------------------------
           Coef.    Std.Err.     z      P>|z|    [0.025    0.975]
-------------------------------------------------------------------
const     -3.9900    1.1400   -3.5001   0.0005   -6.2242   -1.7557
gre        0.0023    0.0011    2.0699   0.0385    0.0001    0.0044
gpa        0.8040    0.3318    2.4231   0.0154    0.1537    1.4544
rank_2    -0.6754    0.3165   -2.1342   0.0328   -1.2958   -0.0551
rank_3    -1.3402    0.3453   -3.8812   0.0001   -2.0170   -0.6634
rank_4    -1.5515    0.4178   -3.7131   0.0002   -2.3704   -0.7325
=================================================================
```

Figure 2: the model of using C-1 variables

Besides, from Figure 2, the coefficients of $rank_1, rank_2, rank_3$ and $rank_4$ have high accuracy (because their Std.Err. are very small), and from the %5 significance level, they are statistically significant.

3 The role of independence

Let X_1 be chosen at random. Then, let dependent Bernoulli r.v. $X_1, ..., X_n(i.e.,$ $X_i \in 0, 1)$ be a sequence of $X_1 = X_2 = ... = X_n$. So, we have

$$\begin{aligned} E(X_i) &= E(X_1) \\ &= 1 * \frac{1}{2} + 0 * \frac{1}{2} = \frac{1}{2} \end{aligned}$$

Then,

$$\begin{aligned} \left| \mu - \frac{1}{n} \sum_{i=1}^{n} X_i \right| &= \left| \frac{1}{2} - 1 \right| \\ &= \frac{1}{2} \end{aligned}$$

So, $P\left(\left| \mu - \frac{1}{n} \sum_{i=1}^{n} X_i \right| \geq \frac{1}{2} \right) = 1$

4 A bound for student's grades

1. Let $\hat{Q} = 100 - \hat{Z}$. Then, the probability of observing $\hat{Z} \leq z$ is

$$\begin{aligned}
P(\hat{Z} \leq z) &= P(100 - \hat{Z} \geq 100 - z) = P(\hat{Q} \geq 100 - z) \\
&\leq \frac{E(\hat{Q})}{100 - z} = \frac{100 - E(\frac{1}{15}\sum_{i=1}^{15} X_i)}{100 - z} \\
&= \frac{100 - p}{100 - z} = \frac{40}{100 - z}
\end{aligned}$$

   Where $100 - \hat{Z} > 0$ and $100 - z > 0$.

   So, $\delta(100 - z_{max}) = 40$, then $z_{max} = -700$.

2. Let $\hat{Q} = 100 - \hat{Z}$. From the Chebyshev's inequality,

$$\begin{aligned}
P(\hat{Z} \leq z) &= P(100 - \hat{Z} \geq 100 - z) = P(\hat{Q} \geq 100 - z) \\
&= P\left(\hat{Q} - E(\hat{Q}) \geq 100 - z - E(\hat{Q})\right) \\
&\leq P\left(\left|\hat{Q} - E(\hat{Q})\right| \geq 60 - z\right) \\
&\leq \frac{Var(\hat{Q})}{(60 - z)^2}
\end{aligned}$$

   Where $60 - z > 0$.

   Because $\hat{Q} \in [0, 100]$, then, we let a random variable $Y \in \{0, 100\}$, $P(Y = 0) = 0.6$, $P(Y = 100) = 0.4$, and $\hat{Y} = \frac{1}{15}\sum_{i=1}^{15} Y_i$. So, $E(\hat{Y}) = 40$, and $Var(\hat{Y}) = \frac{1}{15}(E(Y_1^2) - E^2(Y_1)) = 2400$. From 1, we know that $\mathrm{E}(\hat{Q}) = 40$. Then,

$$\begin{aligned}
P(\hat{Z} \leq z) &\leq P\left(\left|\hat{Q} - E(\hat{Q})\right| \geq 60 - z\right) \\
&\leq \frac{Var(\hat{Q})}{(60 - z)^2} \\
&\leq \frac{Var(\hat{Y})}{(60 - z)^2}
\end{aligned}$$

   So, $\delta(60 - z_{max})^2 = 160$, then $z_{max} \approx 3.43$

3. From the Hoeffding's inequality,

$$\begin{aligned}
P(\hat{Z} \leq z) &= P(E(\hat{Z}) - \hat{Z} \geq E(\hat{Z}) - z) \\
&= P(E(\hat{Z}) - \hat{Z} \geq 60 - z) \\
&\leq e^{-\frac{2*15^2(60-z)^2}{15*100^2}}
\end{aligned}$$

Where $60 - z > 0$.

So, $e^{-\frac{2*15^2(60-z_{max})^2}{15*100^2}} = 0.05$, then, $z_{max} \approx 28.40$

4. The Chebyshev's inequality and the Hoeffding's inequality provide non-vacuous values of z

5 How to split a sample

1. First, we need to prove that $E\left[\hat{L}(\hat{h}^*_{S^{train}}, S^{test})\right] = L(\hat{h}^*_{S^{train}})$:

$$\begin{aligned}
E\left[\hat{L}(\hat{h}^*_{S^{train}}, S^{test})\right] &= E\left[\frac{1}{n^{test}}\sum_{i=i}^{n^{test}} l(\hat{h}^*_{S^{train}}(X_i), Y_i)\right] \\
&= \frac{1}{n^{test}}\sum_{i=i}^{n^{test}} E\left[l(\hat{h}^*_{S^{train}}(X_i), Y_i)\right] \\
&= \frac{1}{n^{test}}\sum_{i=i}^{n^{test}} \hat{L}(\hat{h}^*_{S^{train}}) \\
&= L(\hat{h}^*_{S^{train}})
\end{aligned}$$

Next, from the Hoeffding's inequality, we have:

$$P\left(L(\hat{h}^*_{S^{train}}) - \hat{L}(\hat{h}^*_{S^{train}}, S^{test}) \geq \epsilon\right) \leq e^{-2n^{test}\epsilon^2}$$

And then, let $\delta = e^{-2n^{test}\epsilon^2}$, so, $\epsilon = \sqrt{\frac{\ln\frac{1}{\delta}}{2n^{test}}}$. We have:

$$P\left(L(\hat{h}^*_{S^{train}}) - \hat{L}(\hat{h}^*_{S^{train}}, S^{test}) \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n^{test}}}\right) \leq \delta \iff$$

$$P\left(L(\hat{h}^*_{S^{train}}) - \hat{L}(\hat{h}^*_{S^{train}}, S^{test}) \leq \sqrt{\frac{\ln\frac{1}{\delta}}{2n^{test}}}\right) \geq 1 - \delta \iff$$

$$P\left(L(\hat{h}^*_{S^{train}}) \leq \sqrt{\frac{\ln\frac{1}{\delta}}{2n^{test}}} + \hat{L}(\hat{h}^*_{S^{train}}, S^{test})\right) \geq 1 - \delta$$

Then, we have a bound on $L(\hat{h}^*_{S^{train}})$ in terms of $\hat{L}(\hat{h}^*_{S^{train}}, S^{test})$ and n^{test} that holds with probability at least $1 - \delta$.

2. Let $H = \{\hat{h}_1^*, \hat{h}_2^*, ..., \hat{h}_m^*\}$ and $\delta_i = \frac{\delta}{m}$, then, we have:

$$
\begin{aligned}
&P\left(\forall \hat{h}_i^* \in H : L(\hat{h}_i^*) \geq \sqrt{\frac{\ln \frac{1}{\delta_i}}{2n_i}} + \hat{L}(\hat{h}_i^*, S_i^{test})\right) \\
&\leq P\left(\exists \hat{h}_i^* \in H : L(\hat{h}_i^*) \geq \sqrt{\frac{\ln \frac{1}{\delta_i}}{2n_i}} + \hat{L}(\hat{h}_i^*, S_i^{test})\right) \\
&\leq \sum_{\hat{h}^* \in H} P\left(L(\hat{h}_i^*) \geq \sqrt{\frac{\ln \frac{1}{\delta_i}}{2n_i}} + \hat{L}(\hat{h}_i^*, S_i^{test})\right) \\
&\leq \sum_{\hat{h}^* \in H} \delta_i = \sum_{\hat{h}^* \in H} \frac{\delta}{m} = \delta
\end{aligned}
$$

This equals:

$$
P\left(\forall \hat{h}_i^* \in H : L(\hat{h}_i^*) \geq \sqrt{\frac{\ln \frac{m}{\delta}}{2n_i}} + \hat{L}(\hat{h}_i^*, S_i^{test})\right) \leq \delta \iff
$$

$$
P\left(\forall \hat{h}_i^* \in H : L(\hat{h}_i^*) \leq \sqrt{\frac{\ln \frac{m}{\delta}}{2n_i}} + \hat{L}(\hat{h}_i^*, S_i^{test})\right) \geq 1 - \delta
$$

Where, we can have a fixed $n_i = \frac{1}{2\epsilon^2} \ln \frac{m}{\delta}$, so all the n_i have the same value.

Then, we have a bound on $L(\hat{h}_i^*)$ in terms of $\hat{L}(\hat{h}_i^*, S_i^{test})$ and n_i that holds for all $\hat{h}_i^*$ simultaneously with probability at least $1 - \delta$.