Machine Learning
Assignment 2
Template

Mao Xiqing (tls868)

November 30, 2020

Contents

1 Illustration of Hoeffding's Inequality

Source code for this question can be found in the notebook `MCHInequalities.ipynb`.

1. The Hoeffding's bound on $\mathbb{P}(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha)$ for the same values of α with bias 0.5 and 0.1 are shown on following Figure 1 and Figure 2.

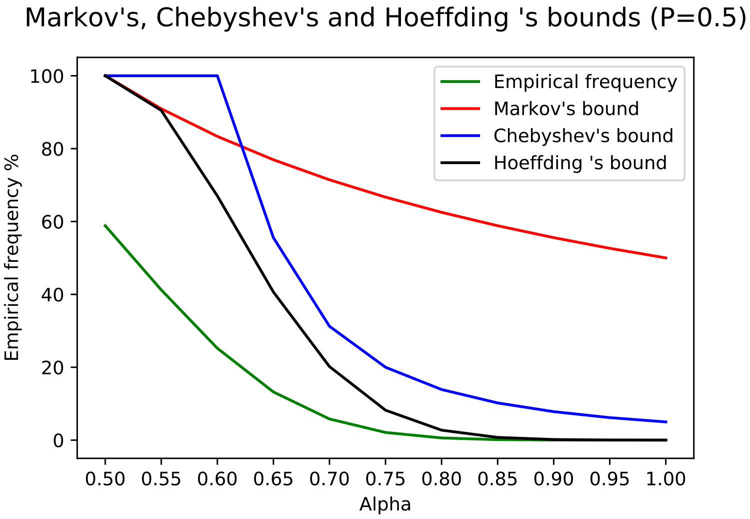

Figure 1: Comparison of Markov's, Chebyshev's and Hoeffding 's bounds with bias 0.5

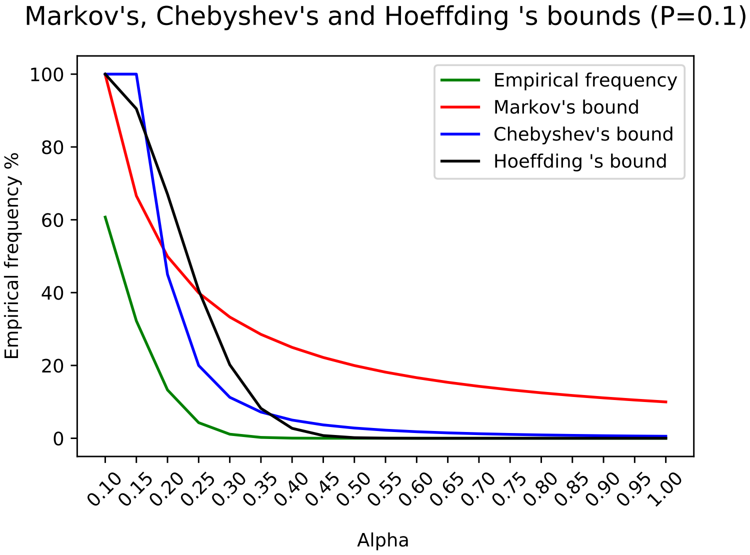

Figure 2: Comparison of Markov's, Chebyshev's and Hoeffding 's bounds with bias 0.1.

Above figures show Hoeffding's inequality gives the tightest bound in most of the cases (except when p close to bias). It bounds the probability that an empirical mean of these random variables deviating from a true mean value, so the bound falls exponentially and more quickly than other two bounds, and it more close to the exact probability. While markov's and Chebyshev's bound focus on the individual sample rather than mean, neither of them is strong enough to show the tight bounds needed for our case, especially when bias is 0.5.

2. Comparison of exact probability and Hoeffding's bound.
   Two Hoeffding's bounds are close to exact probability, both of them are approximate to 0.

Table 1: Comparison of exact probability and Hoeffding's bound with bias 0.5

α Probability	**1**	**0.95**
Exact probability	9.5367×10^{-7}	9.7666×10^{-3}
Hoffding's Inequality	4.5400×10^{-3}	3.0354×10^{-2}

Table 2: Comparison of exact probability and Hoeffding's bound with bias 0.1

α Probability	**1**	**0.95**
Exact probability	1×10^{-20}	1.8×10^{-9}
Hoffding's Inequality	8.4890×10^{-13}	1×2.8111^{-11}

2 The effect of scale (range) and normalization of random variables in Hoeffding's inequality

In **Theorem 2.3**, $X_i \in [a_i, b_i]$:

$$\mathbb{P}(\sum_{n=1}^{N} X_i - \mathbb{E}[\sum_{n=1}^{N} X_i] \geq \epsilon) \leq e^{-2\epsilon^2/\sum_{n=1}^{N}(b_i-a_i)^2}. \tag{1}$$

If $X_i \in [0, 1]$, therefore $a_i = 0, b_i = 0$. Since $\mathbb{E}[\sum_{n=1}^{N} X_i] = np$, each part can be divided by n, then:

$$\mathbb{P}(\frac{1}{n}\sum_{n=1}^{N} X_i - \mathbb{E}[X_i] \geq \epsilon) \leq exp(\frac{-2\epsilon^2 n}{\frac{1}{n}\sum_{n=1}^{N}(1-0)^2}) = e^{-2n\epsilon^2}, \tag{2}$$

and

$$\mathbb{P}(\mathbb{E}[X_i] - \frac{1}{n}\sum_{n=1}^{N} X_i \geq \epsilon) \leq exp(\frac{-2\epsilon^2 n}{\frac{1}{n}\sum_{n=1}^{N}(1-0)^2}) = e^{-2n\epsilon^2}. \tag{3}$$

3 Distribution of Student's Grades

1. Let $\hat{Q} = 100 - \hat{Z}$, from Markov's inequality:

$$\mathbb{P}(\hat{Q} \geq (100 - z)) \leq \frac{\mathbb{E}[\hat{Q}]}{(100 - z)}, \tag{4}$$

where $\mathbb{E}[\hat{Q}] = \mathbb{E}[100] - \mathbb{E}[\hat{Z}]] = 40$ and $100 - z \geq 0$. Then $\delta(100 - z) = 40$, hence $z_{max} = -700$.

2. From Chebyshe's Inequality:

$$\begin{aligned}\mathbb{P}(|\hat{Q} - \mathbb{E}[\hat{Q}]| \geq (100 - z - \mathbb{E}[\hat{Q}])] &= \mathbb{P}(|\hat{Q} - 40| \geq (60 - z)) \\ &\leq \frac{Var[\hat{Q}]}{(100 - z - \mathbb{E}[\hat{Q}])^2} \\ &= \frac{Var[\hat{Q}]}{(60 - z)^2},\end{aligned} \tag{5}$$

where $\hat{Q} \in [0, 100]$ and $60 - z \geq 0$.

Let $\mathbb{P}(\hat{Q} = 0) = p_0$, $\mathbb{P}(\hat{Q} = 100) = 1 - p_0$, then $\mathbb{E}[\hat{Q}] = 0 \times p_0 + 100 \times (1 - p_0) \rightarrow p_0 = 0.6$.

Let Y be a random variable, such that $P(Y = 0 - 40) = 0.6$ and $P(Y = 100 - 40) = 1 - 0.6 = 0.4$, then we have $Var[Q] \leq Var[Y]$:

$$\mathbb{P}(|\hat{Q} - 40| \geq (60 - z)) \leq \frac{Var[\hat{Q}]}{(60 - z)^2} \leq \frac{Var[\hat{Y}]}{(60 - z)^2}, \tag{6}$$

where $Var[Y] = \mathbb{E}[y^2] - (\mathbb{E}[y])^2 = 2400$, then $(2400/7)/(60 - z)^2 \leq \delta = 0.05$, hence $z_{max} = -23$.

3. From Hoeffding's Inequality:

$$\mathbb{P}[(\mathbb{E}[7\hat{Z}] - 7\hat{Z}) \leq (\mathbb{E}[7\hat{Z}] - 7z)] \leq e^{-2(\mathbb{E}[7\hat{Z}] - 7z)^2/(7\times 100^2)}, \tag{7}$$

where $\mathbb{E}[7\hat{Z}] - 7z \geq 0$, then $\delta = 0.05 \geq exp(-2(7(60 - z))^2/(7 \times 100^2)$, hence $z_{max} = 13.749$.

4. Hoeffding's Inequality provide a non-vacuous value of z.

4 The Airline Question

1. Let X be a random variable of the number of no-show people, and $X \sim b(100, 0.05)$. The number of people that show up for a flight will be larger than the number of seats means that $X < (100 - 99) \rightarrow X = 0$.

$$\mathbb{P}(X = 0) = \binom{100}{0} 0.05^0 \times 0.95^{100-0} = 0.5921\% \tag{8}$$

2. In these case, let the first event as A, and let $X_1, ...X_n$ be independent random variables, $X_i \in [0, 1]$, which represent no-show and shows up respectively. To compute $\mathbb{P}(\sum_{n=1}^{10000} X_i = 9500)$, I applied Hoeffding's Inequality:

$$\mathbb{P}(A) = \mathbb{P}(\frac{1}{10000}\sum_{i=1}^{10000} X_i - \mathbb{E}[x_i] \geq 0.95 - \mathbb{E}[x_i]) \leq e^{-2\times 10000\times(0.95-\mathbb{E}[x_i])^2}, \tag{9}$$

where $\mathbb{E}[x_i] = \mathbb{P}(\text{show up})$. And let the second event as B, which is equivalent to the previous question. Since these two events are independent, the probability of them happen simultaneously could be $\mathbb{P} \leq \mathbb{P}(A) \times \mathbb{P}(B)$:

$$\begin{aligned} &\mathbb{P} \leq \mathbb{P}(A) \times \mathbb{P}(B) \\ &= e^{-2\times 10000\times(0.95-p)^2} \times \sum_{seat+1}^{tickets} (\binom{tickets}{seats+1} p^{seats+1} \times (1-p)^{tickets-seats-1}) \\ &= e^{-2\times 10000\times(0.95-p)^2} \times p^{100}. \end{aligned} \tag{10}$$

When $p = 0.952624$, the probability of two events happen simultaneously is 0.00679745. The plot is shown on Figure 3.

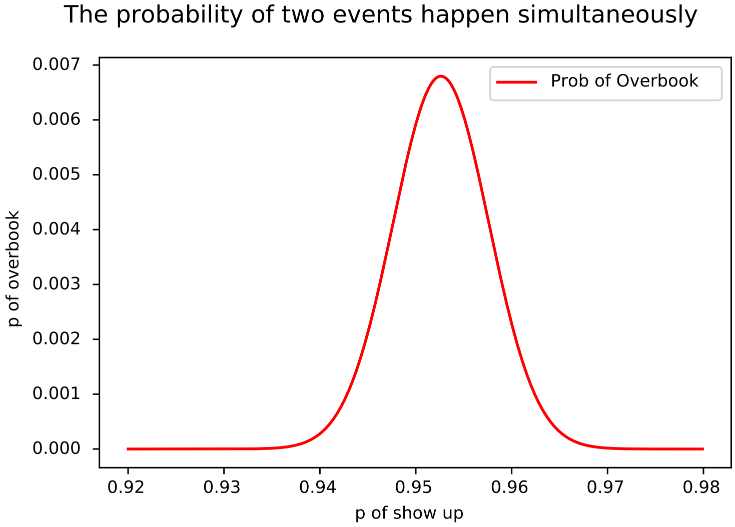

Figure 3: This figure shows the probability of observing 10000 sample and getting a flight overbooked.

5 Linear classification

5.1 Cross-entropy measure

(5.1.a).

According to *Exercise 3.6*, $P(y \mid \boldsymbol{x})$ is the probability of $[\![y_n = \pm 1]\!]$. The likelihood function of it is:

$$\prod_{n=1}^{n} P(y_n \mid \boldsymbol{x_n}), \tag{11}$$

the equivalent method of maximize the likelihood function is to minimize the negative logarithmic likelihood:

$$-\sum_{n=1}^{n} \ln(P(y_n \mid \boldsymbol{x_n})), \tag{12}$$

where

$$P(y_n \mid \boldsymbol{x_n}) = \begin{cases} h(x_n) & \text{for } y = 1 \\ 1 - h(x_n) & \text{for } y = -1. \end{cases} \tag{13}$$

By applying Cross-entropy, the minimizes h can be easily found by the following likelihood function:

$$\begin{aligned} E_{in}(w) &= -\sum_{n=1}^{N} \ln(P(y_n \mid \boldsymbol{x}_n)) \\ &= -\sum_{n=1}^{N} [\![y_n = 1]\!] \ln h(x_n) + [\![y_n = -1]\!] \ln(1 - h(x_n)) \\ &= \sum_{n=1}^{N} [\![y_n = 1]\!] \ln \frac{1}{h(x_n)} + [\![y_n = -1]\!] \ln \frac{1}{1 - h(x_n)}, \end{aligned} \tag{14}$$

(5.1.b).

For the case $h(x) = \theta(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})$:

$$\frac{1}{h(x_n)} = \frac{1}{\theta(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})} = 1/\frac{1}{1 + e^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}} = 1 + e^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}, \tag{15}$$

$$\frac{1}{1 - h(x_n)} = 1 + e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}, \tag{16}$$

therefore

$$\begin{aligned} E_{in}(w) &= \sum_{n=1}^{N} [\![y_n = 1]\!] \ln(1 + e^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}) + [\![y_n = -1]\!] \ln(1 + e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}) \\ &= \sum_{n=1}^{N} \ln(1 + e^{-y_n \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}), \end{aligned} \tag{17}$$

which is equivalent of minimizing the one in equation (3.9).

5.2 Logistic regression loss gradient

When *labels* $\in \{-1, 1\}$,

$$\begin{aligned} \nabla E_{in}(\boldsymbol{w}) &= \frac{\partial}{\partial \boldsymbol{w}}(\frac{1}{N}\sum_{n=1}^{N} \ln(1 + e^{-y_n \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}) \\ &= \frac{1}{N}\sum_{n=1}^{N}[\frac{1}{1 + e^{-y_n \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}} \times e^{-y_n \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}} \times (-y_n \boldsymbol{x})] \\ &= \frac{1}{N}\sum_{n=1}^{N}[\frac{1}{1/(e^{-y_n \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}) + 1} \times (-y_n \boldsymbol{x})] \\ &= -\frac{1}{N}\sum_{n=1}^{N}(\frac{y_n \boldsymbol{x}}{1 + e^{y_n \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}}) \\ &= \frac{1}{N}\sum_{n=1}^{N}[-y_n \boldsymbol{x}\theta(-y_n \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})]. \end{aligned} \tag{18}$$

A misclassified sample, for example, $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} > 0 \rightarrow y = 1$ but classified to $y = -1$, meaning that $-y_n \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} > 0$, and then $\theta(-y_n \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})$ is close to 1, while for a correct classified sample $-y_n \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} < 0$, then $\theta(-y_n \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})$ is close to 0. Therefore a misclassified sample contributes more gradient than correct one.

When *labels* $\in \{0, 1\}$, the likelihood function can also be rewritten this way:

$$\begin{aligned} E_{in}(w) &= -\frac{1}{N}\sum_{n=1}^{N} \ln(P(y_n \mid \boldsymbol{x}_n)) \\ &= -\frac{1}{N}\sum_{n=1}^{N} \ln(p^{y_n}(1-p)^{1-y_n}) \\ &= -\frac{1}{N}\sum_{n=1}^{N}[y_n \ln(p) + (1-y_n)\ln(1-p)], \end{aligned} \tag{19}$$

where $p = \theta(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})$, and

$$\begin{aligned}
p' &= (\frac{1}{1+e^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}})' \\
&= \frac{1}{1+e^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}} \cdot \frac{e^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}}{1+e^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}} \cdot x \\
&= p(1-p)x. \\
(1-p)' &= -p(1-p)x
\end{aligned} \tag{20}$$

Therfore,

$$\begin{aligned}
\nabla E_{in}(\boldsymbol{w}) &= \frac{\partial}{\partial \boldsymbol{w}}\{-\frac{1}{N}\sum_{n=1}^{N}[y_n \ln(p) + (1-y_n)\ln(1-p)]\} \\
&= -\frac{1}{N}\sum_{n=1}^{N}((y_n\frac{1}{p}p') + (1-y_n)\frac{1}{1-p}(1-p)') \\
&= -\frac{1}{N}\sum_{n=1}^{N}(y_n - p)x \\
&= \frac{1}{N}\sum_{n=1}^{N} -[y_n - \theta(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})]x.
\end{aligned} \tag{21}$$

Now it is the same as the one on slide 18.

If $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} > 0$ misclassified, $-(y_n - \theta(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})) = -(0 - \theta(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}))$ is close to 1. But when correct classified, $-(y_n - \theta(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})) = -(1 - \theta(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x})) = -(\theta(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}))$ is close to 0 (because $1-\theta(s) = e^{-s}/(1+e^{-s}) = 1/(1+\frac{1}{e^{-s}}) = 1/(1+e^{s}) = \theta(-s)$), so a misclassified sample contributes more gradient than correct one. In other word, only misclassified one contributes the gradient.

5.3 Log-odds

Prove:
Assuming that

$$P(Y = y \mid X = \boldsymbol{x}) = \begin{cases} p & \text{for } y = 1 \\ 1-p & \text{for } y = 0, \end{cases} \tag{22}$$

then

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b = ln\frac{P(Y=1 \mid X=\boldsymbol{x})}{P(Y=0 \mid X=\boldsymbol{x})} = ln\frac{p}{1-p}. \tag{23}$$

Take e for both sides:

$$\begin{aligned}
e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b} &= e^{ln\frac{p}{1-p}} \\
&= \frac{p}{1-p},
\end{aligned} \tag{24}$$

then

$$P(Y=1 \mid X=\boldsymbol{x})=p=\frac{1}{1+e^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}, \tag{25}$$

then σ is the logistic function:

$$f(\mathbf{x})=\sigma\left(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b\right)=\frac{1}{1+e^{-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b}}=P(Y=1 \mid X=\boldsymbol{x}). \tag{26}$$

5.4 Variable importance

Source code for this question can be found in the notebook `one-hot.ipynb`. Here is the result of using one-hot encoding:

```
Optimization terminated successfully.
         Current function value: 0.573147  Iterations 6
                            Results: Logit
=======================================================================
......
-----------------------------------------------------------------------
        Coef.    Std.Err.       z     P>|z|      [0.025          0.975]
-----------------------------------------------------------------------
const  -3.9054 8488674.4571 -0.0000 1.0000 -16637500.1177 16637492.3069
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 8488674.4571 -0.0000 1.0000 -16637496.2969 16637496.1277
rank_2 -0.7600 8488674.4571 -0.0000 1.0000 -16637496.9723 16637495.4523
rank_3 -1.4248 8488674.4571 -0.0000 1.0000 -16637497.6371 16637494.7875
rank_4 -1.6360 8488674.4571 -0.0000 1.0000 -16637497.8484 16637494.5763
=======================================================================
```

There are two main reasons that we should use dummy variables.

Firstly, it is enough to use C-1 variables instead of C variables, because if no ranking is 2,3 and 4, the ranking must be 1, so it can be (0,0,0) instead of complicated (1,0,0,0).

Besides, if using one-hot encoding, collinearity would be introduced. The model of using it would be

$$y=w_1 GRE+w_2 GPA+D_1 rank_1+D_2 rank_2+D_3 rank_3+D_4 rank_4+b, \tag{27}$$

where $rank_1+rank_2+rank_3+rank_4=1$, and therefore there must exist $rank_1+rank_2+rank_3+rank_4-1=0$, which means there are infinite solutions for optimization problem, which also means the coefficient may not be true and variance increase (8488674), as well as significance become meaningless. That is why it would be difficult to interpret the variable importance if we only use one-hot encoding .