ML-hw1

February 2019

Group member:
Cheng Shen cs3750
Qidong Yang qy2216
Jerry Lin sl4299

Problem1

$\frac{df(x+tu)}{dt}|_{t=0}$

$= \lim_{t\to 0} \frac{f(x+tu)-f(x)}{t}$

$\nabla f(x) = \lim_{t\to 0} \frac{f(x+tu)-f(x)}{tu}$

$u \cdot \nabla f(x) = \lim_{t\to 0} \frac{f(x+tu)-f(x)}{t}$

Thus

$\frac{df(x+tu)}{dt}|_{t=0} = \nabla f(x) \cdot u$

$\nabla f(x) \cdot u$

$\leq \|\nabla f(x)\| \cdot \|u\|$

$= \frac{\nabla f(x)}{\|\nabla f(x)\|} \cdot \nabla f(x)$

$= \nabla f(x) \cdot v$

Accroding to Cauchy–Schwarz inequality, only when u=v, the equality can be achieved. As a reuslt,

$\frac{df(x+tu)}{dt}|_{t=0} \leq \frac{df(x+tv)}{dt}|_{t=0}$

Problem 2

2.1

Given the original error function Eq. (1) and the new error function Eq. (2)

$$\min_f \mathbb{E}_{(X,Y)}[\mathbf{1}[f(X) \neq Y]], \tag{1}$$

$$\min_f \mathbb{E}_{(X,Y)}[\max\{0, 1 - Yf(X)\}], \tag{2}$$

we can generate a truth table as follows.

f(X)	Y	$\mathbf{1}$[f(X) $\neq Y$]	$\max\{0, 1 - Yf(X)\}$)
0	0	0	0
0	1	1	2
1	0	1	2
1	1	0	0

It is clear that when the classifier f makes a correct prediction, (i.e. $f(X) = Y$), we have in Eq. (2)

$$\max\{0, 1 - Yf(X)\} = 0,$$

which is the same behavior as that of $\mathbf{1}[f(X) \neq Y]$ in Eq. (1). Furthermore, false positive and false negative predictions (i.e. $f(X) \neq Y$) yield maximum positive results for $\max\{0, 1 - Yf(X)\}$ in Eq. (2), which is also the same behavior for $\mathbf{1}[f(X) \neq Y]$ in Eq. (1).

2.2

$$\mathbb{E}_{(X,Y)}[\max\{0, 1 - Yf(X)\}] \tag{3}$$

$$= \mathbb{E}_{(X,Y)}[2 \cdot \max\{0, .5 - .5 \cdot Yf(X)\}] \tag{4}$$

$$= \mathbb{E}_{(X,Y)}[2 \cdot \mathbf{1}[f(X) \neq Y] \tag{5}$$

$$=2 \cdot \mathbb{E}_{(X,Y)}[\mathbf{1}[f(X) \neq Y] \tag{6}$$

$$=2 \cdot P_{(X,Y)}[f(X) \neq Y] \tag{7}$$

$$=2 \cdot P[f(x) \neq Y | X = x] \tag{8}$$

$$=2 \cdot (P[f(x) = 0, Y = 1 | X = x] + P[f(x) = 1, Y = 0 | X = x]) \tag{9}$$

$$=2 \cdot (\mathbf{1}[f(x) = 0] \cdot P[Y = 1 | X = x] + \mathbf{1}[f(x) = 1] \cdot P[Y = 0 | X = x]) \tag{10}$$

2.3

The relationship that needs to hold is $Pr(Y = 1|X = x) > Pr(Y = -1|X = x)$. Given certain features x, an optimal binary classifier chooses the most likely outcome seen in the training data set. Therefore, in this case, if an optimal classifier chooses 1 given some features x, this means that the probability of $Y = 1$ given those features, i.e. $Pr(Y = 1|X = x)$, must be greater than $Pr(Y = -1|X = x)$.

2.4

2.4.1

In general, binary classifier chooses the class with higher probability seen in the training set given some features. An optimal classifier can take the form

$$f^*(x) = sign(Pr[Y = 1|X = x] - \frac{1}{2}) \tag{11}$$

because of the following reasoning. Since it is clear that

$$Pr(Y = 1|X = x) + Pr(Y = -1|X = x) = 1,$$

In the case of $Pr(Y = 1|X = x) > Pr(Y = -1|X = x)$, we have

$$Pr[Y = 1|X = x] - \frac{1}{2} \in [0, \frac{1}{2}],$$

which is further attributed to 1 by the *sign* function. On the other hand, in the case of $Pr(Y = -1|X = x) > Pr(Y = 1|X = x)$, we have

$$Pr[Y = 1|X = x] - \frac{1}{2} \in [-\frac{1}{2}, 0],$$

which is then scaled to -1 by the *sign* function. Therefore, Eq. (11) well captures the desired behavior of an optimal classifier.

2.4.2

The classifier, shown as Eq. (11), is closely related to the Optimal Bayes Classifier. We can draw the following equations:

$$\begin{aligned} f^*(x) &= sign(Pr[Y=1|X=x]-\frac{1}{2}) && (12) \\ &= \arg\max_{Y\in\{-1,1\}}(Pr[Y=y|X=x]-\frac{1}{2}) && (13) \\ &= \arg\max_{Y\in\{-1,1\}}(Pr[Y=y|X=x]). && (14) \end{aligned}$$

We can see that Eq. (12) and Eq. (13) share the same behavior of choosing the class with higher probability. Eq. (12) does so by checking whether $Pr[Y=1|X=x]$ is greater or less than $\frac{1}{2}$ and generating corresponding output. Eq. (13) does so by returning the maximizing argument to $Pr[Y=y|X=x]$. Then Eq. (13) derives Eq. (14), which is the definition of a binary Bayes Classifier.

2.5

We can rewrite the expectation as the following:

$$\begin{aligned} &\mathbb{E}_{(X,Y)}[(1-Yf(X))^2] \\ =&Pr(Y=1|X=x)\cdot(1-(1)\cdot f(x))^2+Pr(Y=-1|X=x)\cdot(1-(-1)\cdot f(x))^2 \\ =&Pr(Y=1|X=x)\cdot(1-f(x))^2+Pr(Y=-1|X=x)\cdot(1+f(x))^2 \\ =&Pr(Y=1|X=x)\cdot(1-f(x))^2+(1-Pr(Y=1|X=x))\cdot(1+f(x))^2 \\ =&Pr(Y=1|X=x)\cdot(1-f(x))^2+(1+f(x))^2-Pr(Y=1|X=x)\cdot(1+f(x))^2 \\ =&Pr(Y=1|X=x)\cdot((1-f(x))^2-(1+f(x))^2)+(1+f(x))^2 \\ =&Pr(Y=1|X=x)\cdot((1-f(x)+1+f(x))(1-f(x)-1-f(x)))+(1+f(x))^2 \\ =&f^2(x)-4\cdot Pr(Y=1|X=x)\cdot f(x)+2\cdot f(x)+1 \end{aligned}$$

Then we can take the derivative of the expected value with respect to $f(x)$ and set it to 0 to find the minimum.

$$\frac{d\,\mathbb{E}}{df(x)}=2\cdot f(x)-4\cdot Pr(Y=1|X=x)+2=0$$

Therefore we have

$$\begin{aligned} f(x) =& \frac{4 \cdot Pr(Y = 1|X = x) - 2}{2} && (15) \\ =& 2 \cdot Pr(Y = 1|X = x) - 1 && (16) \end{aligned}$$

We can verify that the above classifier, Eq. (16), minimizes the expectation by testing it under specific cases. When $Pr(Y = 1|X = x) = 0$, $f(x) = -1$ and $(1 - Yf(X))^2$ equals 0, which indicates that the error is 0. So we have our optimal classifier

$$f^*(x) = 2 \cdot Pr(Y = 1|X = x) - 1.$$

Problem 3

3.1

$$p(x) = (2\pi)^{-d/2}\exp\{-||x-\mu||^2/2\} = (2\pi)^{-d/2}\exp\{-\frac{1}{2}(x-\mu)^T(x-\mu)\} \tag{1}$$

$p\,(x)$ is proportional to $\log\,p\,(x)$.

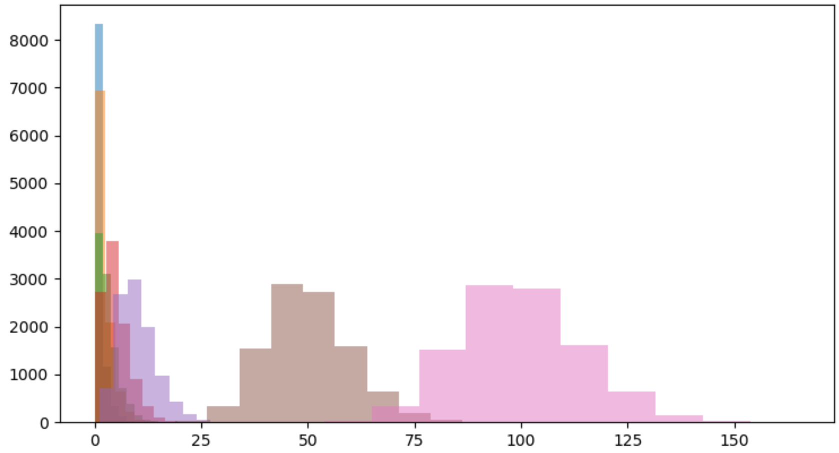
$$\begin{aligned}\log\;p(x) &= \log\;(2\pi)^{-d/2} + \log\;\exp\{-\frac{1}{2}(x-\mu)^T(x-\mu)\}\\ &= -\frac{d}{2}\log(2\pi) - \frac{1}{2}(x-\mu)^T(x-\mu)\\ &= -\frac{d}{2}\log(2\pi) - \frac{1}{2}(x^Tx - 2x^T\mu + \mu^T\mu)\end{aligned} \tag{2}$$

Then take the derivative with respect to x.

$$\begin{aligned}\frac{d\log\;p(x)}{dx} &= -\frac{1}{2}(2x-2\mu)\\ &= -x+\mu = 0\end{aligned} \tag{3}$$

The above formula implies that $x = \mu$ is the only critical point. And the Hessian matrix of $\log p\,(x)$ is $-I$, which is always negative definite. Therefore, $\log p\,(x)$ is concave. The only critical point μ is global max.

3.2



The centers of different dimension's histograms are running away from origin. Higher dimension, larger histogram center value. Most samples do not lie close to the mean, especially for high dimension cases.

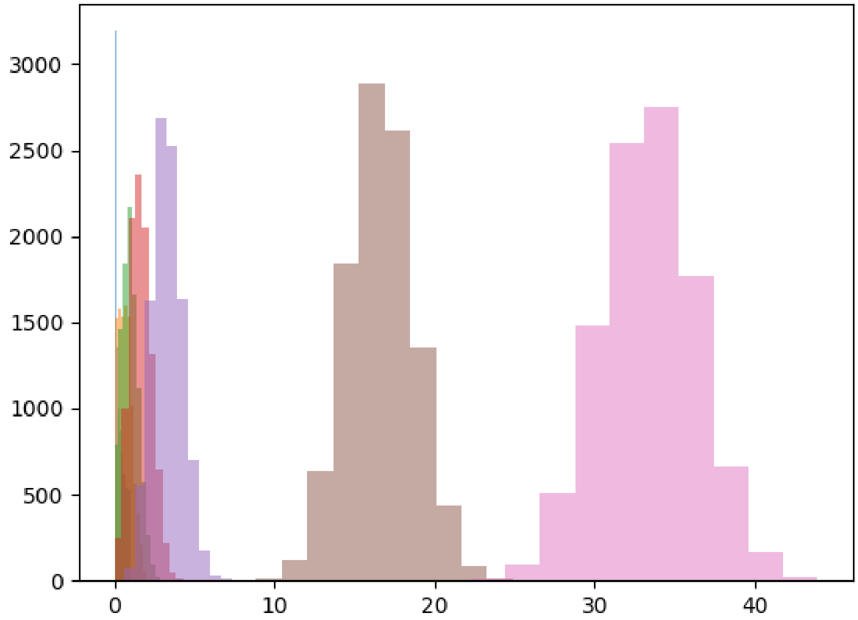
3.3

Because $x \sim N\left(0, I_d\right)$, every component x_i from x is subject to the normal distribution $N\left(0,1\right)$ independently. Var(x_i) =E(x_i^2) −E$(x_i)^2 = 1$. E$(x_i)^2 = 0$. E$(x_i^2) = 1$.

$$\begin{aligned}\mathbb{E}_{x\sim N(0,I_d)}[||x||^2] &= \mathbb{E}_{x\sim N(0,I_d)}[x_1^2 + \cdots + x_i^2 + \cdots + x_d^2] \\ &= \mathbb{E}_{x\sim N(0,I_d)}[\sum_{i=1}^{d} x_i^2] \\ &= \sum_{i=1}^{d} \mathbb{E}_{x\sim N(0,I_d)}[x_i^2] \\ &= \sum_{i=1}^{d} 1 \\ &= d\end{aligned} \tag{4}$$

The plot agrees with the math expression.

3.4



3.5

Because x ~unif$\big([-1,1]^d\big)$, every component x_i from x is also subject to the uniform distribution unif($[-1,1]$). Var(x_i) =E(x_i^2) −E$(x_i)^2 = 1/3$. E$(x_i)^2 = 0$. E$(x_i^2) = 1/3$.

$$\begin{aligned}
\mathbb{E}_{x\sim \text{unif}([-1,1]^d)}[||x||^2] &= \mathbb{E}_{x\sim \text{unif}([-1,1]^d)}[x_1^2 + \cdots + x_i^2 + \cdots + x_d^2] \\
&= \mathbb{E}_{x\sim \text{unif}([-1,1]^d)}[\sum_{i=1}^{d} x_i^2] \\
&= \sum_{i=1}^{d} \mathbb{E}_{x\sim \text{unif}([-1,1]^d)}[x_i^2] \\
&= \sum_{i=1}^{d} \frac{1}{3} \\
&= \frac{1}{3}d
\end{aligned} \tag{5}$$

The plot agrees with the math expression.

Problem 4

4.1

$$\begin{aligned}\frac{\Pr[s=1|x]}{\Pr[s=1|y=1]} &= \underbrace{\frac{\Pr[s=1,y=1|x]}{\Pr[s=1|y=1]}}_{\text{s=1 implies y=1}} \\ &= \underbrace{\frac{\Pr[s=1,y=1|x]}{\Pr[s=1|y=1,x]}}_{\text{s,x conditionally independent}} \\ &= \frac{\frac{\Pr[s=1,y=1,x]}{\Pr[x]}}{\frac{\Pr[s=1,y=1,x]}{\Pr[y=1,x]}} \\ &= \frac{\Pr[y=1,x]}{\Pr[x]} \\ &= \Pr[y=1|x]\end{aligned} \tag{6}$$

4.2

$$\begin{aligned}\frac{\Pr[s=1|x]}{1-\Pr[s=1|x]} &= \underbrace{\frac{\Pr[s=1,y=1|x]}{1-\Pr[s=1,y=1|x]}}_{\text{s=1 implies y=1}} \\ &= \frac{\Pr[s=1,y=1,x]}{\Pr[x]-\Pr[s=1,y=1,x]}\end{aligned} \tag{7}$$

$$\begin{aligned}\frac{1-\Pr[s=1|y=1]}{\Pr[s=1|y=1]} &= \underbrace{\frac{1-\Pr[s=1|y=1,x]}{\Pr[s=1|y=1,x]}}_{\text{s,x conditionally independent}} \\ &= \frac{\Pr[y=1,x]-\Pr[s=1,y=1,x]}{\Pr[s=1,y=1,x]}\end{aligned} \tag{8}$$

$$\begin{aligned}RHS &= \frac{\Pr[y=1,x]-\Pr[s=1,y=1,x]}{\Pr[s=1,y=1,x]}\frac{\Pr[s=1,y=1,x]}{\Pr[x]-\Pr[s=1,y=1,x]} \\ &= \frac{\Pr[y=1,x]-\Pr[s=1,y=1,x]}{\Pr[x]-\Pr[s=1,y=1,x]} \\ &= \frac{\Pr[s=0,y=1,x]}{\Pr[x]-\Pr[s=1,y=1,x]} \\ &= \underbrace{\frac{\Pr[s=0,y=1,x]}{\Pr[s=0,x]}}_{\text{total probability rule}} \\ &= \Pr[y=1|x,s=0]\end{aligned} \tag{9}$$

Note the last second step requires that $\Pr[s=1|x,y=0]=0$ and $\Pr[s=1,y=0,x]=0$, which is provided.

4.3

$$\begin{aligned}
\Pr[f(x) \neq y|x] &= \Pr[y=0|x]\mathbf{1}[f(x) \neq 0] + \Pr[y=1|x]\mathbf{1}[f(x) \neq 1] \\
&= \Pr[y=0, s=0|x]\mathbf{1}[f(x) \neq 0] + \Pr[y=0, s=1|x]\mathbf{1}[f(x) \neq 0] \\
&+ \Pr[y=1, s=0|x]\mathbf{1}[f(x) \neq 1] + \Pr[y=1, s=1|x]\mathbf{1}[f(x) \neq 1] \\
&= \Pr[y=0|s=0,x]\Pr[s=0|x]\mathbf{1}[f(x) \neq 0] \\
&+ \Pr[y=0|s=1,x]\Pr[s=1|x]\mathbf{1}[f(x) \neq 0] \\
&+ \Pr[y=1|s=0,x]\Pr[s=0|x]\mathbf{1}[f(x) \neq 1] \\
&+ \Pr[y=1|s=1,x]\Pr[s=1|x]\mathbf{1}[f(x) \neq 1]
\end{aligned} \tag{10}$$

Because s=1 implies y=1, $\Pr[y=1|s=1,x] = 1$ and $\Pr[y=0|s=1,x] = 0$.

$$\begin{aligned}
\Pr[f(x) \neq y|x] &= \Pr[y=0|s=0,x]\Pr[s=0|x]\mathbf{1}[f(x) \neq 0] \\
&+ \Pr[y=0|s=1,x]\Pr[s=1|x]\mathbf{1}[f(x) \neq 0] \\
&+ \Pr[y=1|s=0,x]\Pr[s=0|x]\mathbf{1}[f(x) \neq 1] \\
&+ \Pr[y=1|s=1,x]\Pr[s=1|x]\mathbf{1}[f(x) \neq 1] \\
&= \Pr[y=0|s=0,x]\Pr[s=0|x]\mathbf{1}[f(x) \neq 0] \\
&+ \Pr[y=1|s=0,x]\Pr[s=0|x]\mathbf{1}[f(x) \neq 1] \\
&+ \Pr[s=1|x]\mathbf{1}[f(x) \neq 1]
\end{aligned} \tag{11}$$

$$\begin{aligned}
\mathbb{E}_{(x,y)\sim D}[\mathbf{1}[f(x) \neq y]] &= \Pr[f(x) \neq y] \\
&= \int_x \Pr[f(x) \neq y|x]\Pr[x]dx \\
&= \int_x \Pr[y=0|s=0,x]\Pr[s=0,x]\mathbf{1}[f(x) \neq 0] \\
&+ \Pr[y=1|s=0,x]\Pr[s=0,x]\mathbf{1}[f(x) \neq 1] \\
&+ \Pr[s=1,x]\mathbf{1}[f(x) \neq 1]dx \\
&= \int_x \Pr[x,s=1]\mathbf{1}[f(x) \neq 1] \\
&+ \Pr[x,s=0](\Pr[y=1|s=0,x]\mathbf{1}[f(x) \neq 1] \\
&+ \Pr[y=0|s=0,x]\mathbf{1}[f(x) \neq 0])dx
\end{aligned} \tag{12}$$

4.4

When sample is selected, we use following to estimate.

$$\frac{1}{N_1}\sum_{i:s_i=1}^{N_1} \mathbf{1}[f(x_i) \neq 1] \tag{13}$$

When sample is not selected, we use following to estimate.

$$\frac{1}{N_2}\sum_{i:s_i=0}^{N_2} \mathbf{1}[f(x_i) \neq 1]w(x_i) + \mathbf{1}[f(x_i) \neq 0](1 - w(x_i)) \tag{14}$$

In particular, from part 2, we know that $\Pr[y = 1|x, s = 0]$ and $\Pr[y = 0|x, s = 0]$ can be estimated from (x, s) only. Following hint , we represent them as $w\,(x)$ and $1 - w\,(x)$.

In conclusion, the final answer is:

$$\begin{aligned}
&\frac{1}{N_1}\sum_{i:s_i=1}^{N_1} \mathbf{1}[f(x_i) \neq 1] \\
&+ \frac{1}{N_2}\sum_{i:s_i=0}^{N_2} \mathbf{1}[f(x_i) \neq 1]w(x_i) + \mathbf{1}[f(x_i) \neq 0](1 - w(x_i))
\end{aligned} \tag{15}$$

Problem 5

5.1

Because other attributes may have correlation with the sensitive attribute. For example, when we deal with the crime data, the race is a sensitive data. When we just remove the race attribute, other attributes like salary, education background which have correlation with the race attribute still cause unfairness. For example, the high salary group may have low crime rate.

5.2

$P[\hat{Y}=1] = P_a[\hat{Y}=1] \quad \forall a \in \{0,1\}$

When a = 1,

$P_1[\hat{Y}=1] = P_1[\hat{Y}=1]$

$P_0[\hat{Y}=1] = P_1[\hat{Y}=1]$

When a = 0,

$P_1[\hat{Y}=1] = P_0[\hat{Y}=1]$

$P_0[\hat{Y}=1] = P_0[\hat{Y}=1]$

Thus, we can get $P_0[\hat{Y}=1] = P_0[\hat{Y}=1]$ from the four equations above. This proves the equivalence.

5.3

$P_{a_1}[\hat{Y}=k] = P_{a_2}[\hat{Y}=k] \iff P[\hat{Y}=k] = P_a[\hat{Y}=k] \ \forall a, a_1, a_2 \in N, \forall k \in R$

5.4

The code is submitted to Gradescope.

5.5

We tested on different sizes of training data. First, we trained on original training data. In this round, we ran the knn classifier for different k and norm options. Second, we trained on smaller size of training data. The two size options are 1000 and 2000. As we can see from the figures, MLE classifier is the best calssifier among all the classifiers on all the training data with different sizes.

The training sample size has impacts on performance on test data. We used the original train data set, sample of size 2000 and sample of size 1000. We can see from the plot that the test accuracy will go down when we decrease the sample size.

5.6

In this part, we choose $y = 0 and \hat{y} = 0$. We calculated the probability when a=0 and a=1. Then, we use the absolute value of the corresponding probability difference as

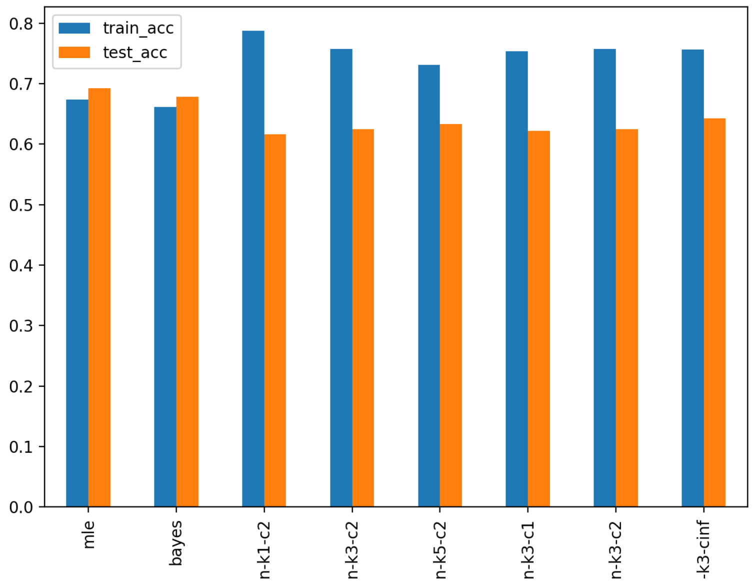
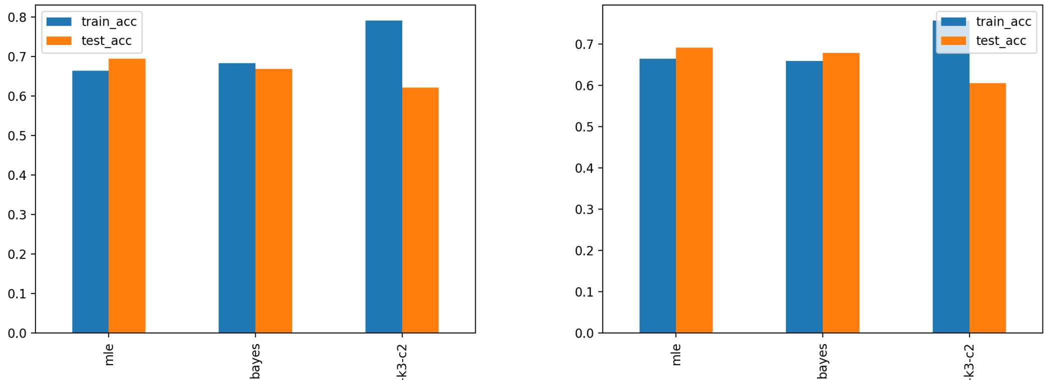
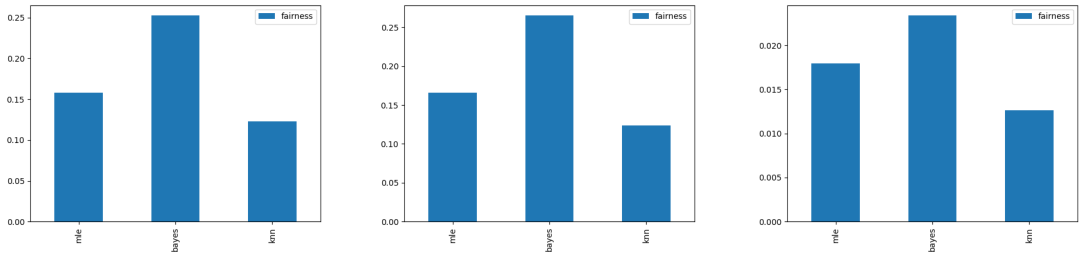


Figure 1: Original train data size, set different for knn classifier



(a) Training data size is 1000

(b) Training data size is 2000



(a) Fairness difference in DP

(b) Fairness difference in EO

(c) Fairness difference in PP

the metric to evaluate the degree of the fairness definition. As we can see, difference in demographic parity and the difference in equalized odds are around 0.15 for all the three classifiers. The fairness is not good. For predictive parity, the differences are around 0.015 which means the difference is very small for all the three classifier. In other words, the three classifier achieves predictive parity in this case.

Moreover, knn classifier has best performance in all three definitions. Bayes classifier has the worst performance.

5.7

I choose Demographic Parity. Suppose $Y \in \{0, 1\}$ refers to whether a student has GPA greater than 3.5, $A \in K$ refers to the student's ethnic. If we want the ethnic attribute to be insensitive and fair. Then, students from different ethnic should have the same probability to achieve GPA over 3.5. Otherwise, the attribute is unfair. This is reasonable.

One potential disadvantage is that if we force such fairness requirement, we may miss the best classifier. Sometimes, the base rate in training data is different. For example, certain group of students are hardworking, the difference in base rate may lead to miss the best classifier.