crwc

COMS 4771 Machine Learning (Spring 2018)
Problem Set #1

Zhuangyu Ren, Yuchen Mo, Chengzhi Mao - UNI: zr2209, ym2706, cm3797

February 25, 2019

Problem 1

Proof. Let $p = x + tu$ and with the chain rule we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}u} f(x+tu) &= \frac{\mathrm{d}f(p)}{\mathrm{d}p} \cdot \frac{\mathrm{d}p}{\mathrm{d}u} \\ &= \nabla f(x) \cdot u \\ &\leq ||u|| \cdot ||\nabla f(x)|| \end{aligned} \tag{1}$$

$\frac{\mathrm{d}}{\mathrm{d}u} f(x+tu) = ||u|| \cdot ||\nabla f(x)||$ iff $\vec{u}$ is parallel to $\nabla f(x)$. And at that time we have

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

Which is exactly the $\vec{v}$ we want. And with this knowledge we can also transform (1) to

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}u} f(x+tu) &\leq ||u|| \cdot ||\nabla f(x)|| \\ &= \frac{\mathrm{d}}{\mathrm{d}v} f(x+tv)\end{aligned}$$

□

Problem 2

2.1

The original function E is to minimize the number of cases that $f(X) \neq Y$
Here in this new function, we consider two cases:
if $f(X) \neq Y$, then $Yf(x) = -1$, $\max\{0, 1 - Yf(x)\} = 2$
if $f(X) = Y$, then $Yf(x) = 0$, $\max\{0, 1 - Yf(x)\} = 0$
Here we want to minimize E, which is $\mathbb{E}_{(X,Y)}[\max\{0, 1 - Yf(x)\}]$. So we should minimize $\max\{0, 1 - Yf(x)\}$, and its minimum value is 0, at where $f(X) = Y$.
This means that we should minimize the number of cases that $f(X) \neq Y$, so it is similar to the original equation.

2.2

From problem 2.1 we can rewrite the expectation in the following form:

$$\begin{aligned}\mathbb{E}_{(X,Y)}[\max\{0, 1 - Yf(x)\}] &= \mathbb{E}[\mathbf{1}[f(X) \neq Y]] \\ &= \Pr[f(x) \neq y | X = x] \\ &= \Pr[f(x) = 1, y = -1 | X = x] + \Pr[f(x) = -1, y = 1 | X = x] \\ &= \mathbf{1}[f(x) = 1] \cdot \Pr[y = -1 | X = x] + \mathbf{1}[f(x) = -1] \cdot \Pr[y = 1 | X = x]\end{aligned}$$

2.3

The expectation

$$\mathbf{1}[f(x) = 1] \cdot \Pr[y = -1 | X = x] + \mathbf{1}[f(x) = -1] \cdot \Pr[y = 1 | X = x]$$

can take two values:
If $f(x) = 1$, it equals to $\Pr[y = -1 | X = x]$,
If $f(x) = -1$, it equals to $\Pr[y = 1 | X = x]$.
As we want our optimal classifier $f^*(x) = 1$ to minimize the expectation, it means that

$$\Pr[Y = 1 | X = x] > \Pr[Y = -1 | X = x]$$

always holds to make the expectation value minimal.

2.4

For the optimal Bayes Classifier, the optimal function is

$$f^*_{\text{Bayes}}(x) = \arg\max_{y \in \{-1,1\}} \Pr[Y = y | x = x]$$

Here we have

$$\begin{aligned} f^*(x) &= \operatorname{sign}\left(\Pr\left[Y=1|X=x\right] - \frac{1}{2}\right) \\ &= \operatorname{sign}\left(\frac{1}{2}\Pr\left[Y=1|X=x\right] + \frac{1}{2}\left(1 - \Pr\left[Y=-1|X=x\right]\right) - \frac{1}{2}\right) \\ &= \operatorname{sign}\left(\Pr\left[Y=1|X=x\right] - \Pr\left[Y=-1|X=x\right]\right) \\ &= \arg\max_{y\in\{-1,1\}} \Pr\left[Y=y|X=x\right] \end{aligned}$$

Which is exactly the same as the optimal function for Bayes Classifier.

2.5

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

We want to minimize E over all f, which is the same as minimizing

$$f(x)^2 + (2-4\Pr[Y=1|X=x])f(x) + 1$$

over all f.
It is easy to see that the minimizer is at

$$f(x) = -\frac{2-4\Pr[Y=1|X=x]}{2}$$

So the optimal classifier is

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

Problem 3

3.1

Proof. It's easy to see that for any $x \in \mathbb{R}^d$

$$p(\mu) - p(x) = (2\pi)^{-d/2}\left(1 - e^{-||x-\mu||^2/2}\right)$$

As $0 \leq e^{-||x-\mu||^2/2} \leq 1,\ \forall x \in \mathbb{R}^d,\ p(\mu) - p(x) \geq 0.$ □

3.2

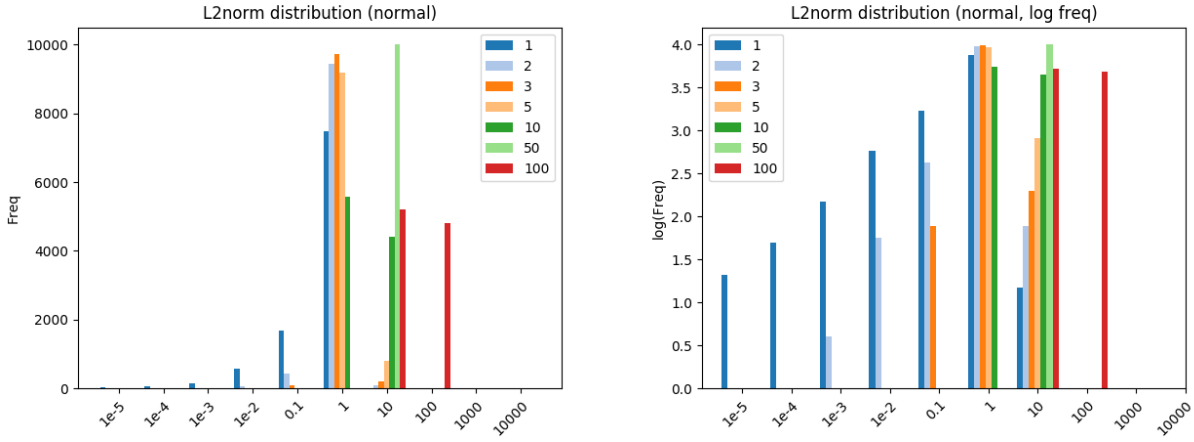

Figure 1: L2norm distribution by normal distribution (original and log scaled)

It's not easy to display the L2 norm without using log scale, and the drawback of this method is we cannot see the mean of those samples directly from the figure. In this case, our program printed the mean value of each experiment while plotting.
1 mean: 1.0096564193179323
2 mean: 2.004663734524429
3 mean: 3.0051783025853616
5 mean: 5.028034439987408
10 mean: 10.008850413964852
50 mean: 50.03963153085344
100 mean: 99.86147438303487
It's observed that when the dimension goes high, the samples' L2 norm is about proportional to the number of dimensions. Most of samples are close to the mean of their group, but not the mean of the original Gaussian distribution's mean 0.

3.3

As the covariance matrix is diagonal, each dimension of x can be considered as sampled from i.i.d. 1D standard normal distribution. Thus let's denote $||x||^2$ as t. We'll have

$$t = ||x||^2 = x_1^2 + \cdots + x_d^2$$

Therefore, t satisfies chi-squared distribution with d dof. $\mathbb{E}[t] = d$. This result matches our previous observation.
Alternative solution:

$$\begin{aligned}\mathbb{E}[||x||^2] &= \sum_i \mathbb{E}[x_i^2] \\ &= \sum_i \left(\mathbb{E}[x_i]^2 + \mathrm{Var}(x_i)\right) \\ &= d(\mu^2 + 1) = d\end{aligned}$$

3.4

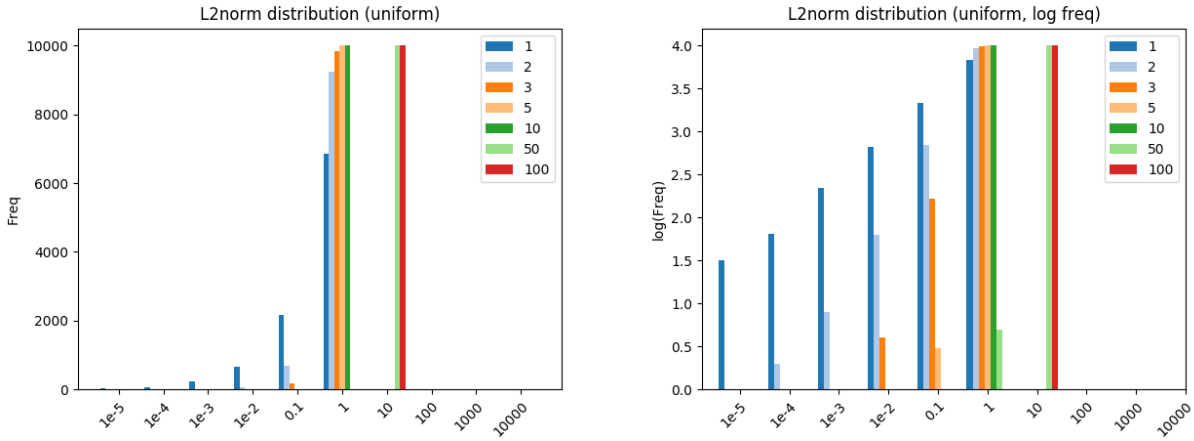

Figure 2: L2norm distribution by uniform distribution (original and log scaled)

And the mean value of each group of samples are:
1 mean: 0.3315518420816778
2 mean: 0.6724232688900863
3 mean: 1.000128888676283
5 mean: 1.6650075946017535
10 mean: 3.3223475677670873
50 mean: 16.662545359727606
100 mean: 33.36863752225879

This time it's observed that the mean is about 1/3 of the number of dimensions.

3.5

$$\begin{aligned}\mathbb{E}_{x\sim\text{unif}([-1,1]^d)}\left[||x||^2\right] &= d\cdot\mathbb{E}_{x\sim\text{unif}([-1,1])}\left[||x||^2\right]\\ &= d\cdot\int_{-1}^{1}x^2\cdot\frac{1}{2}\mathrm{d}x\\ &= \frac{d}{3}\end{aligned}$$

Which is also in agreement with our previous observation.
Alternative solution:

$$\begin{aligned}\mathbb{E}[||x||^2] &= \sum_i\left(\mathbb{E}[x_i]^2+\text{Var}(x_i)\right)\\ &= d(0+\frac{1}{3}) = \frac{d}{3}\end{aligned}$$

Problem 4

4.1

Proof. From the problem we know that:

$$\Pr[s = 1|x, y = 0] = 0$$

So using Bayes formula we have:

$$\Pr[s = 1|x, y = 0] = \frac{\Pr[s = 1] \cdot \Pr[y = 0|x, s = 1]}{\Pr[y = 0|x]} = 0$$

Because $\Pr[y = 0|x]$ and $\Pr[s = 1]$ should not equal to 0, so

$$\Pr[y = 0|x, s = 1] = 0$$

Thus $\Pr[y = 1|x, s = 1] = 1$
By Bayes:

$$\Pr[y = 1|x, s = 1] = \frac{\Pr[y = 1|x] \cdot \Pr[s = 1|x, y = 1]}{\Pr[s = 1|x]} = 1$$

As given y, s and x are conditionally independent, $\Pr[s = 1|x, y = 1] = \Pr[s = 1|y = 1]$
So,

$$\frac{\Pr[y = 1|x] \cdot \Pr[s = 1|y = 1]}{\Pr[s = 1|x]} = 1$$

$$\Pr[y = 1|x] = \frac{\Pr[s = 1|x]}{\Pr[s = 1|y = 1]}$$

Alternative solution:

$$\Pr[y = 1|x] \cdot \Pr[s = 1|y = 1] = \Pr[s = 1, y = 1|x]$$

Also we have

$$\Pr[s = 1|x, y = 1] = \frac{\Pr[s = 1, y = 0|x]}{\Pr[s = 1|x]} = 0 \Rightarrow \Pr[s = 1, y = 0|x] = 0$$

So

$$\Pr[y = 1|x] \cdot \Pr[s = 1|y = 1] = \Pr[s = 1, y = 1|x] + \Pr[s = 1, y = 0|x] = \Pr[s = 1|x]$$

$$\Rightarrow \Pr[y = 1|x] = \frac{\Pr[s = 1|x]}{\Pr[s = 1|y = 1]}$$

□

4.2

From problem 1 we know that

$$\Pr[y=1|x,s=0] = \frac{\Pr[y=1|x]\cdot\Pr[s=0|x,y=1]}{\Pr[s=0|x]} = \frac{\Pr[s=1|x]}{\Pr[s=1|y=1]}\cdot\frac{\Pr[s=0|x,y=1]}{\Pr[s=0|x]}$$

Also

$$\Pr[s=0|x] = 1-\Pr[s=1|x]$$
$$\Pr[s=0|x,y=1] = 1-\Pr[s=1|x,y=1]$$

As given y, s and x are conditionally independent, $\Pr[s=1|x,y=1] = \Pr[s=1|y=1]$
So the origin formula can be written as:

$$\Pr[y=1|x,s=0] = \frac{1-\Pr[s=1|y=1]}{\Pr[s=1|y=1]}\cdot\frac{\Pr[s=1|x]}{1-\Pr[s=1|x]}$$

4.3

$$\begin{aligned}
\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbf{1}[f(x)\neq y]] &= \int_x \Pr[f(x)\neq y]\mathrm{d}x \\
&= \int_x \Pr[f(x)=1,y=0] + \Pr[f(x)=0,y=1]\mathrm{d}x \\
&= \int_x \Pr[f(x)=0]\cdot\Pr[y=1|x] + \Pr[f(x)=1]\cdot\Pr[y=0|x]\mathrm{d}x \\
&= \int_x \Pr[f(x)=0]\left(\Pr[y=1,s=0|x]+\Pr[y=1,s=1|x]\right) \\
&\qquad + \Pr[f(x)=1](\Pr[y=0,s=0|x]+\Pr[y=0,s=1|x])\mathrm{d}x \\
&= \int_x \mathbf{1}[f(x)\neq 1]\left(\Pr[y=1,s=0|x]+\Pr[y=1,s=1|x]\right) \\
&\qquad + \mathbf{1}[f(x)\neq 0]\cdot\Pr[y=0,s=0|x]\mathrm{d}x
\end{aligned}$$

We have proved that

$$\Pr[y=1|x,s=1] = \frac{\Pr[y=1,s=1|x]}{p(x,s=1)} = 1$$

So

$$\Pr[y=1,s=1|x] = p(x,s=1)$$
$$\begin{aligned}
\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbf{1}[f(x)\neq y]] = \int_x &\mathbf{1}[f(x)\neq 1]\cdot p(x,s=1) + \mathbf{1}[f(x)\neq 1]\cdot\Pr[y=1,s=0|x] \\
&+\mathbf{1}[f(x)\neq 0]\cdot\Pr[y=0,s=0|x]\mathrm{d}x
\end{aligned}$$

Because

$$\Pr[y=1,s=0|x] = p(x,s=0)\cdot\Pr[y=1|x,s=0]$$

and

$$\Pr[y = 0, s = 0|x] = p(x, s = 0) \cdot \Pr[y = 0|x, s = 0]$$

So

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbf{1}[f(x) \neq y]] = \int_x & \mathbf{1}[f(x) \neq 1] \cdot p(x, s = 1) \\ & + p(x, s = 0)(\Pr[y = 1|s = 0, x] \cdot \mathbf{1}[f(x) \neq 1] \\ & \quad + \Pr[y = 0|s = 0, x] \cdot \mathbf{1}[f(x) \neq 0])\mathrm{d}x\end{aligned}$$

4.4

Under the assumption that there exists $x \in X$ such that $\Pr[Y = 1|X = x] = 1$ then $\max_{x\in X} g(x) = \Pr[S = 1|Y = 1]$ where $g(x) = \Pr[S = 1|X = x]$. Note that g(x) (and hence its max) can be estimated from (x, s) data only.
Alternatively, $\Pr[S = 1|Y = 1]$ is just a single number (does not depend on X)
So suppose $c = \Pr[s = 1|y = 1]$, Here, c is the constant probability that a positive example is labeled.
Let $w(x) = \Pr[y = 1|x, s = 0]$,

$$\begin{aligned}\Pr[y = 1|x, s = 0] &= \frac{1 - \Pr[s = 1|y = 1]}{\Pr[s = 1|y = 1]} \cdot \frac{\Pr[s = 1|x]}{1 - \Pr[s = 1|x]} \\ &= \frac{1 - c}{c} \cdot \frac{\Pr[s = 1|x]}{1 - \Pr[s = 1|x]}\end{aligned}$$

And the estimator can be written as:

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbf{1}[f(x) \neq y]] &= \int_x \mathbf{1}[f(x) \neq 1] \cdot p(x, s = 1) \\ &\qquad + p(x, s = 0)(\Pr[y = 1|s = 0, x] \cdot \mathbf{1}[f(x) \neq 1] \\ &\qquad\quad + \Pr[y = 0|s = 0, x] \cdot \mathbf{1}[f(x) \neq 0])\mathrm{d}x \\ &= \int_x \mathbf{1}[f(x) \neq 1] \cdot p(x, s = 1) \\ &\qquad + p(x, s = 0)(w(x) \cdot \mathbf{1}[f(x) \neq 1] + (1 - w(x))\mathbf{1}[f(x) \neq 0])\mathrm{d}x\end{aligned}$$

As the number of samples is limited,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbf{1}[f(x) \neq y]] = \frac{1}{|S|}\left[\sum_{x,s=1} \mathbf{1}[f(x) \neq 1] + \sum_{x,s=0} (w(x) \cdot \mathbf{1}[f(x) \neq 1] + (1 - w(x))\mathbf{1}[f(x) \neq 0])\right]$$

Problem 5

5.1

The sensitive attribute A might not be completely independent with other attributes. For example, attribute A is race or religion and there is another attribute B indicates the nationality. Due to the high relevance among these attributes, we can still get some information of the 'sensitive attribute A' from attribute B.
A more bruteforce example is that attribute B is set to be 2A. Then removing A from the dataset obviously means nothing.

5.2

As $\mathbb{P}[\hat{Y} = 1] = \mathbb{P}_\alpha[\hat{Y} = 1]$ holds true for all $\alpha \in \{0, 1\}$, it means that the following two equations all hold:

$$\mathbb{P}[\hat{Y} = 1] = \mathbb{P}_0[\hat{Y} = 1]$$
$$\mathbb{P}[\hat{Y} = 1] = \mathbb{P}_1[\hat{Y} = 1]$$

So

$$\mathbb{P}[\hat{Y} = 1] = \mathbb{P}_0[\hat{Y} = 1] = \mathbb{P}_1[\hat{Y} = 1]$$

If we suppose the ratio of A=0 in the dataset is λ and the ratio of A=1 is $1 - \lambda$, P can be represented as:

$$\mathbb{P}[\hat{Y} = 1] = \lambda\mathbb{P}_0[\hat{Y} = 1] + (1 - \lambda)\mathbb{P}_1[\hat{Y} = 1]$$

Because $\mathbb{P}_0[\hat{Y} = 1] = \mathbb{P}_1[\hat{Y} = 1]$, so:

$$\mathbb{P}[\hat{Y} = 1] = \mathbb{P}_0[\hat{Y} = 1] = \mathbb{P}_1[\hat{Y} = 1]$$

The two definitions for Demographic Parity are equivalent.

5.3

$$\mathbb{P}\alpha_1[\hat{Y} = \hat{y}] = \mathbb{P}\alpha_2[\hat{Y} = \hat{y}] \Leftrightarrow \mathbb{P}[\hat{Y} = \hat{y}] = \mathbb{P}\alpha[\hat{Y} = \hat{y}] \qquad \forall \hat{y} \in \mathbb{R},\ \forall \alpha_1, \alpha_2, \alpha \in \mathbb{N}$$

5.4

Code submitted via Courseworks.

5.5

We tested these three models with different size of training dataset. By sampling from our current dataset, we can plot the relationship between training set size and test accuracy. In each experiment, we increase our size of input dataset from 5% of the original size to 100% with step size 5% (except that we didn't experiment 5% data on KNN because in that case data size is smaller than k). For each parameter, we do 10 experiments and take the mean accuracy.

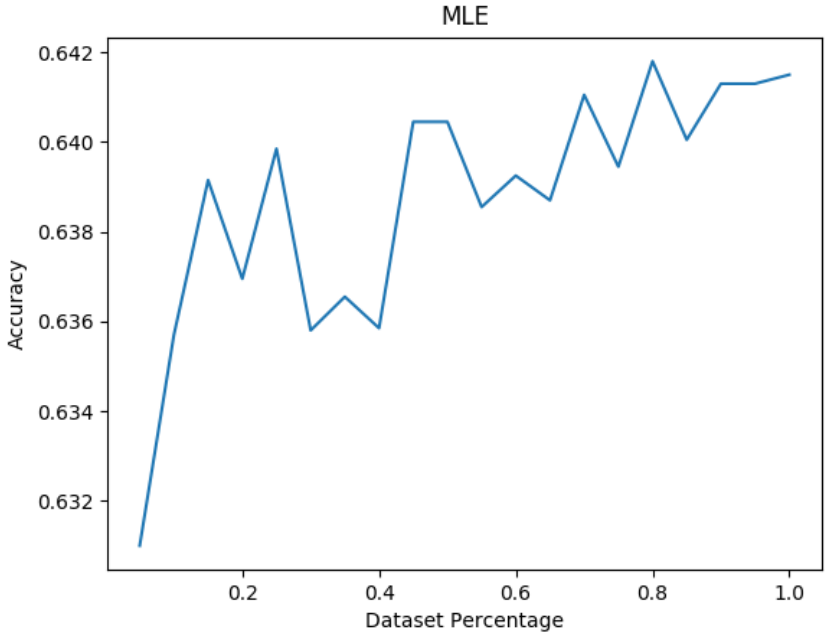

Figure 3: MLE

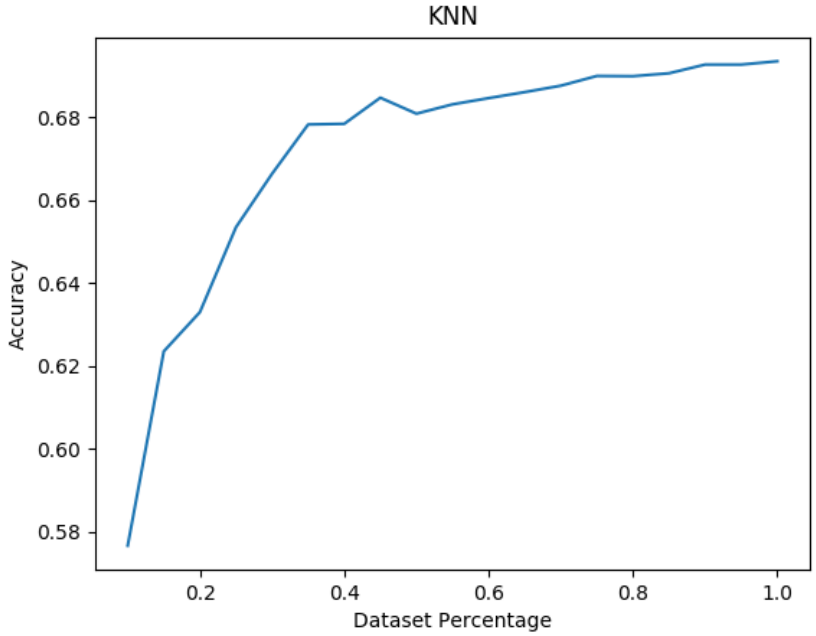

Figure 4: KNN

It can be inferred that when the dataset is small, MLE and Naive Bayes method performs better than KNN. Intuitively, if the dataset size is comparable with the k in KNN then this algorithm obviously won't work well. Therefore, model based Bayes method could achieve

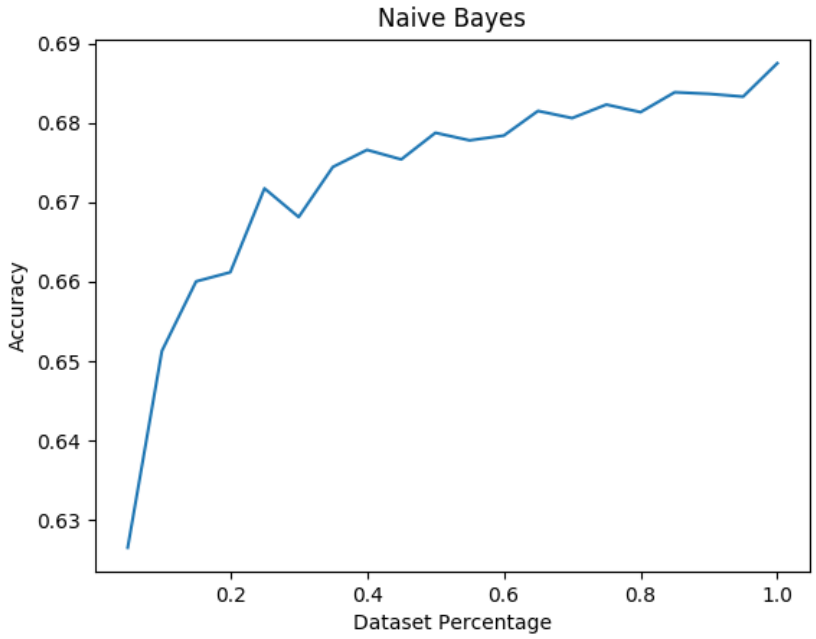

Figure 5: Naive Bayes

better results with their prior information. However, as the training data size grows, it's hard for MLE to increase its accuracy while KNN and Naive Bayes are still working well. This time MLE's idea to estimate a certain model's parameters restricted its possibility of improvement.
And between KNN and Naive Bayes I think here Naive Bayes is a better choice for this prediction task. First of all, KNN's running time complexity is proportional to the training data size, making it harder to evaluate when our dataset grows larger. Then, KNN needs to select a k carefully, while there is no need to manually set a parameter for Naive Bayes.
Even though, I think these conclusions are not so reliable as the performance of these models can vary due to different pre-processing implementations, like how to normalize the inputs.

5.6

Fairness	Model	$\hat{y}$=0		$\hat{y}$=1	
		y=0	y=1	y=0	y=1
DP	$\mathbb{P}_0$	0.469		0.1646	
	$\mathbb{P}_1$	0.5798		0.0768	
EO	$\mathbb{P}_0$	0	0.7896	0	0.2103
	$\mathbb{P}_1$	0	0.8961	0	0.1039
PP	$\mathbb{P}_0$	0.5943	0.4056	0.2171	0.7829
	$\mathbb{P}_1$	0.6471	0.3530	0.2610	0.7391

Table 1: Fairness experiments (MLE)

Fairness	Model	$\hat{y}$=0		$\hat{y}$=1	
		y=0	y=1	y=0	y=1
DP	$\mathbb{P}_0$	0.3982		0.2941	
	$\mathbb{P}_1$	0.5271		0.1687	
EO	$\mathbb{P}_0$	0	0.5891	0	0.4109
	$\mathbb{P}_1$	0	0.7515	0	0.2485
PP	$\mathbb{P}_0$	0.6760	0.3240	0.2841	0.7158
	$\mathbb{P}_1$	0.7014	0.2986	0.3212	0.6787

Table 2: Fairness experiments (KNN)

Fairness	Model	$\hat{y}$=0		$\hat{y}$=1	
		y=0	y=1	y=0	y=1
DP	$\mathbb{P}_0$	0.3675		0.3226	
	$\mathbb{P}_1$	0.5406		0.1415	
EO	$\mathbb{P}_0$	0	0.5299	0	0.4701
	$\mathbb{P}_1$	0	0.7924	0	0.2078
PP	$\mathbb{P}_0$	0.6935	0.3035	0.3137	0.6863
	$\mathbb{P}_1$	0.6825	0.3174	0.3188	0.6811

Table 3: Fairness experiments (Naive Bayes)

These tables showed all the probabilities used to calculate those measurements. According to the graph, by computing the relative ratio, we can see that the KNN has the best DP and EO fairness, the Naive Bayes classifier is most fair with respect to PP fairness.

5.7

For Demographic Parity:
We can not predict whether the man is a criminal/terrorist by his race. So whatever the race is, we should always predict the same result. However, DP requires that other attributes that has correlation to race, such as nationality, won't contribute to this prediction result, although they are not so sensitive. This will led to a loss in prediction accuracy in exchange for fairness in some perspective.