Instructor:
Prof. Nakul Verma
COMS W4771–Spring, 2022

Student:
Vishweshwar Tyagi
vt2353 @columbia.edu

Homework 1

Due: 11:59pm, February 18, 2022

Homework Problems

Solution-1:

(i) We have $x_i \stackrel{iid}{\sim} \text{Bin}(1, b)$. Hence, the likelihood is given by

$$\begin{aligned} L(b) &= \prod_{i=1}^{n} p(x_i|b) \\ &= \prod_{i=1}^{n} b^{x_i}(1-b)^{1-x_i} \\ &= b^{\sum\limits_i x_i}(1-b)^{n-\sum\limits_i x_i} \end{aligned}$$

The log-likelihood is given by

$$LL(b) = \ln b \cdot \sum_i x_i + \ln(1-b) \cdot \big(n - \sum_i x_i\big)$$

We now find the MLE estimate for b, which is $\hat{b}_{MLE}$

$$\begin{aligned} & LL'(b) = 0 \\ \rightarrow\ & \frac{1}{b} \cdot \sum_i x_i - \frac{1}{1-b} \cdot \big(n - \sum_i x_i\big) = 0 \\ \rightarrow\ & (1-b) \cdot \sum_i x_i = b \cdot \big(n - \sum_i x_i\big) \\ \rightarrow\ & \hat{b}_{MLE} = \bar{x} \end{aligned}$$

Let us verify that this indeed is a point of maximum

$$LL''(b) = -\frac{1}{b^2} \cdot \sum_i x_i - \frac{1}{(1-b)^2} \cdot \big(n - \sum_i x_i\big)$$

which is less than 0 since $\sum\limits_i x_i \leq n$ and either $\sum\limits_i x_i > 0$ or $\big(n - \sum\limits_i x_i\big) > 0$

(ii) We know that $\sum\limits_{i=1}^{n} x_i \sim \text{Bin}(n, b)$. Hence, $\mathbb{E}\, \hat{b}_{MLE} = \mathbb{E}\, \bar{x} = \frac{1}{n} \cdot \mathbb{E}\, \sum\limits_{i=1}^{n} x_i = \frac{1}{n} \cdot nb = b$.

Therefore, the MLE estimator is unbiased.

Now, we know that $\mathbb{V}\, \hat{b}_{MLE} = \mathbb{V}\ \bar{x} = \frac{1}{n^2} \cdot \mathbb{V}\, \sum\limits_{i=1}^{n} x_i = \frac{nb(1-b)}{n^2} \to 0$ as $n \to \infty$

So, the MLE is unbiased and its variance approaches 0 as the number of samples tend to ∞. Therefore, it is consistent too.

(iii) We know that $x_i \sim \text{Bin}(1, b)$. So, $\mathbb{E}\, {x_i}^2 = P(x_i = 1) = b$

Now, $\mathbb{V}\, x_i = E\, {x_i}^2 - (E\, x_i)^2 = b - b^2 = b(1-b)$. Hence, the required variance is $b(1-b)$

(iv) We know that MLE estimators are invariant under any mapping. Since, $\bar{x}$ is the MLE estimator of b, therefore, $\bar{x}(1-\bar{x})$ is the MLE estimator of the variance, which is $b(1-b)$ a function of b.

(v) We have $P(b|\mathbf{x}) \propto P(\mathbf{x}|b) \cdot p(b)$. Further, we are given that $p(b) \propto 2b$ for $b \in [0,1]$, and $p(b) = 0$ otherwise.

Hence,

$$\begin{aligned}
\underset{b}{\text{argmax}}\ P(b|\mathbf{x}) &= \underset{b}{\text{argmax}}\ P(\mathbf{x}|b) \cdot p(b) \\
&= \underset{b}{\text{argmax}}\ \ln\big(P(\mathbf{x}|b) \cdot p(b)\big) \\
&= \underset{b}{\text{argmax}}\ \ln P(\mathbf{x}|b) + \ln p(b) \\
&= \underset{b}{\text{argmax}}\ LL(b) + \ln p(b) \\
&= \underset{b}{\text{argmax}}\ \ln b \cdot \sum_i x_i + \ln(1-b) \cdot \big(n - \sum_i x_i\big) + \ln(2b)
\end{aligned}$$

Let $f(b) = \ln b \cdot \sum\limits_i x_i + \ln(1-b) \cdot \big(n - \sum\limits_i x_i\big) + \ln(2b)$. Consider,

$$\begin{aligned}
& f'(b) = 0 \\
\to\ & \frac{1}{b} \cdot \sum_i x_i - \frac{1}{1-b} \cdot \big(n - \sum_i x_i\big) + \frac{1}{b} = 0 \\
\to\ & (1-b) \cdot \big(\sum_i x_i + 1\big) = b \cdot \big(n - \sum_i x_i\big) \\
\to\ & \hat{b}_{MAP} = \frac{\sum\limits_{i=1}^{n} x_i + 1}{n+1}
\end{aligned}$$

is the required MAP estimate.

We can verify that it indeed maximizes the objective,

$$f''(b) = -\frac{1}{b^2} \cdot \sum_i x_i - \frac{1}{(1-b)^2} \cdot \big(n - \sum_i x_i\big) - \frac{1}{b^2} < 0$$

(vi) Note that, $\hat{b}_{MLE} = \underset{b}{\text{argmax}}\, LL(b)$, whereas $\hat{b}_{MAP} = \underset{b}{\text{argmax}}\, LL(b) + \ln p(b)$

Clearly, MAP estimate will equal MLE estimate if the given prior $p(b)$ is the uniform distribution. MLE estimate can be thought of as a special case of MAP estimate where the prior is uniform.

Solution-2:

(i) We have the profit π given by

$$\pi = \begin{cases} (P-C)Q & D \geq Q \\ (P-C)D - C(Q-D) & D < Q \end{cases}$$

So, the expected profit is given by

$$\begin{aligned}
\mathbb{E}_{P(D)}\, \pi &= \int_{D\geq Q} (P-C)Q \cdot f(D)\, dD + \int_{0\leq D\leq Q} \big((P-C)D - C(Q-D)\big) \cdot f(D)\, dD \\
&= (P-C)Q \cdot \big(1 - F(Q)\big) + (P-C) \cdot \int_{D\leq Q} Df(D)\, dD \\
&\quad - CQ \int_{D\leq Q} f(D)\, dD + C \int_{D\leq Q} Df(D)\, dD \\
&= (P-C)Q \cdot \big(1 - F(Q)\big) + P \cdot \int_{D\leq Q} Df(D)\, dD - CQF(Q) \\
&= (P-C)Q - PQF(Q) + CQF(Q) + P \cdot \int_{D\leq Q} Df(D)\, dD - CQF(Q) \\
&= (P-C)Q - PQF(Q) + P \cdot \int_{D\leq Q} Df(D)\, dD
\end{aligned}$$

(ii) Let $f(Q) = \mathbb{E}_{P(D)}\, \pi = (P-C)Q - PQF(Q) + P \cdot \int_{D\leq Q} Df(D)\, dD$. Consider

$$\begin{aligned}
& f'(Q) = 0 \\
& \to P - C - PF(Q) - PQf(Q) + PQf(Q) = 0 \\
& \to F(Q) = 1 - \frac{C}{P}
\end{aligned}$$

Hence, the optimal Q^* satisfies $F(Q^*) = 1 - \frac{C}{P}$, as required.

Note that $f''(Q) = -Pf(Q) < 0$, hence, this Q is optimal.

Solution-3:

(i) Note that $\Pr(g(X) \neq Y) = \int_X \Pr(g(X) \neq Y|X) \Pr(X)\, dX$. Now consider,

$$\begin{aligned}
\Pr(g(X) \neq Y|X) &= 1 - \Pr(g(X) = Y|X) \\
&= 1 - \mathbb{I}_{\{g(X)=1\}} \Pr(Y = 1|X) - \mathbb{I}_{\{g(X)=0\}} \Pr(Y = 0|X) \\
&= 1 - \mathbb{I}_{\{g(X)=1\}} \Pr(Y = 1|X) - \mathbb{I}_{\{g(X)=0\}} (1 - \Pr(Y = 1|X)) \\
&= 1 - \mathbb{I}_{\{g(X)=1\}} \eta(X) - \mathbb{I}_{\{g(X)=0\}} (1 - \eta(X))
\end{aligned}$$

We known that $\mathbb{I}_{\{g(X)=1\}} \iff \eta(X) \geq \frac{1}{2}$. Hence, we have,

$$\begin{aligned}
\mathrm{ERR}(g) &= \Pr(g(X) \neq Y) \\
&= \int_X \Pr(g(X) \neq Y|X) \Pr(X)\, dX \\
&= \int_X \left\{1 - \mathbb{I}_{\{g(X)=1\}} \eta(X) - \mathbb{I}_{\{g(X)=0\}} (1 - \eta(X))\right\} \Pr(X)\, dX \\
&= \int_X \Pr(X)\, dX - \int_X \mathbb{I}_{\{g(X)=1\}} \eta(X) \Pr(X)\, dX - \int_X \mathbb{I}_{\{g(X)=0\}} (1 - \eta(X)) \Pr(X)\, dX \\
&= 1 - \int_{X:\eta(X) \geq 1/2} \eta(X) \Pr(X)\, dX - \int_{X:\eta(X) < 1/2} (1 - \eta(X)) \Pr(X)\, dX \\
&= 1 - \frac{1}{2} \cdot \int_{X:\eta(X) \geq 1/2} (2\eta(X) - 1 + 1) \Pr(X)\, dX - \frac{1}{2} \cdot \int_{X:\eta(X) < 1/2} (1 + 1 - 2\eta(X)) \Pr(X)\, dX \\
&= 1 - \frac{1}{2} \cdot \int_{X:\eta(X) \geq 1/2} (2\eta(X) - 1) \Pr(X)\, dX - \frac{1}{2} \cdot \int_{X:\eta(X) \geq 1/2} \Pr(X)\, dX \\
&\quad - \frac{1}{2} \cdot \int_{X:\eta(X) < 1/2} \Pr(X)\, dX - \frac{1}{2} \cdot \int_{X:\eta(X) < 1/2} (1 - 2\eta(X)) \Pr(X)\, dX \\
&= 1 - \frac{1}{2} \cdot \int_X \Pr(X)\, dX - \frac{1}{2} \cdot \int_X \left|2\eta(X) - 1\right| \Pr(X)\, dX \\
&= \frac{1}{2} - \frac{1}{2} \cdot \mathbb{E}_X \left|2\eta(X) - 1\right|
\end{aligned}$$

(ii) Again note that $\Pr(g(X) \neq Y) = \int_X \Pr(g(X) \neq Y|X) \Pr(X)\, dX$. Now consider,

$$
\begin{aligned}
\Pr(g(X) \neq Y|X) &= 1 - \Pr(g(X) = Y|X) \\
&= 1 - \mathbb{I}_{\{g(X)=1\}} \Pr(Y=1|X) - \mathbb{I}_{\{g(X)=0\}} \Pr(Y=0|X) \\
&= 1 - \mathbb{I}_{\{g(X)=1\}} \Pr(Y=1|X) - \mathbb{I}_{\{g(X)=0\}} (1 - \Pr(Y=1|X)) \\
&= \eta(X) + (1-\eta(X)) - \mathbb{I}_{\{g(X)=1\}} \eta(X) - \mathbb{I}_{\{g(X)=0\}}(1-\eta(X)) \\
&= \eta(X)(1 - \mathbb{I}_{\{g(X)=1\}}) + (1-\eta(X))(1 - \mathbb{I}_{\{g(X)=0\}}) \\
&= \mathbb{I}_{\{g(X)=0\}} \eta(X) + \mathbb{I}_{\{g(X)=1\}}(1-\eta(X)) \\
&= \begin{cases} \eta(X) & \text{if } \Pr(Y=0|X) \geq \Pr(Y=1|X) \\ 1-\eta(X) & \text{if } \Pr(Y=0|X) < \Pr(Y=1|X) \end{cases} \\
&= \begin{cases} \Pr(Y=1|X) & \text{if } \Pr(Y=0|X) \geq \Pr(Y=1|X) \\ \Pr(Y=0|X) & \text{if } \Pr(Y=0|X) < \Pr(Y=1|X) \end{cases} \\
&= \min(\Pr(Y=1|X),\ \Pr(Y=0|X)) \\
&= \min\Big(\frac{\Pr(X|Y=1)\Pr(Y=1)}{\Pr(X)},\ \frac{\Pr(X|Y=0)\Pr(Y=0)}{\Pr(X)}\Big) \qquad (1)
\end{aligned}
$$

Therefore, using (1), we get,

$$
\begin{aligned}
\Pr(g(X) \neq Y) &= \int_X \Pr(g(X) \neq Y|X) \Pr(X)\, dX \\
&= \int_X \min\Big(\frac{\Pr(X|Y=1)\Pr(Y=1)}{\Pr(X)},\ \frac{\Pr(X|Y=0)\Pr(Y=0)}{\Pr(X)}\Big) \Pr(X)\, dX \\
&= \int_X \min\big(\Pr(X|Y=1)\Pr(Y=1),\ \Pr(X|Y=0)\Pr(Y=0)\big)\, dX \\
&= \int_X \min\big(p_1 f_1(X),\ (1-p_1) f_0(X)\big)\, dX
\end{aligned}
$$

(iii) Consider,

$$
\begin{aligned}
\mathrm{ERR}(g) &= \frac{1}{2} - \frac{1}{2} \cdot \mathbb{E}_X \big|2\eta(X) - 1\big| \\
&= \frac{1}{2} - \frac{1}{2} \cdot \mathbb{E}_X \big|\Pr(Y=1|X) - (1 - \Pr(Y=1|X))\big| \\
&= \frac{1}{2} - \frac{1}{2} \int_X \big|\Pr(Y=1|X) - \Pr(Y=0|X)\big| \Pr(X)\, dX \\
&= \frac{1}{2} - \frac{1}{2} \int_X \Big|\frac{\Pr(X|Y=1)\Pr(Y=1)}{\Pr(X)} - \frac{\Pr(X|Y=0)\Pr(Y=0)}{\Pr(X)}\Big| \Pr(X)\, dX \\
&= \frac{1}{2} - \frac{1}{2} \int_X \Big|\frac{1}{2} \cdot \Pr(X|Y=1) - \frac{1}{2} \cdot \Pr(X|Y=0)\Big| \cdot \frac{\Pr(X)}{\big|\Pr(X)\big|}\, dX \\
&= \frac{1}{2} - \frac{1}{4} \int_X \big|\Pr(X|Y=1) - \Pr(X|Y=0)\big|\, dX \\
&= \frac{1}{2} - \frac{1}{4} \int_X \big|f_1(X) - f_0(X)\big|\, dX
\end{aligned}
$$

Solution-4:

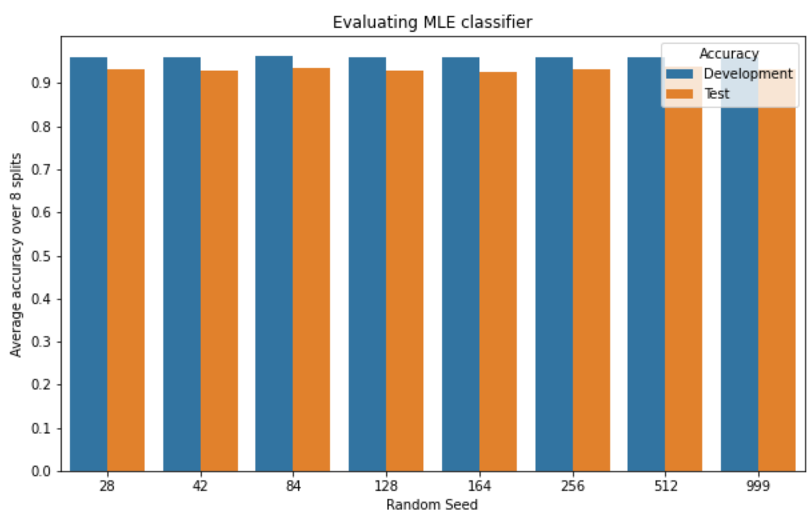
We followed the following structure:

For hyper-parameter tuning: We first split the dataset into development and test set. The development set was further split into K folds. Among these K folds, the model was trained on $K-1$ folds (called training data), and tested on the remaining fold (called validation data). This gave a total of K train and validation scores and the hyper-parameter that gave the highest average validation score was chosen to be optimal. After choosing this optimal hyper-parameter, the model was then trained on all of the development set (all of the K folds) and finally tested on the test dataset.

For model evaluation: We made K splits of the dataset into development and test data. We then trained the model on the development set and tested it on the test set. This gave K development and test accuracy scores.

For model comparison: We compare the performance of MLE and KNN classifiers as the size of training data varies. For a given size s of training data, we consider 5 development and test splits and calculate the average development and test scores for both models.

(i) We evaluated the MLE classifier on 8 different development-test splits in the ratio of 8 : 2 by varying the random seed while creating the splits.

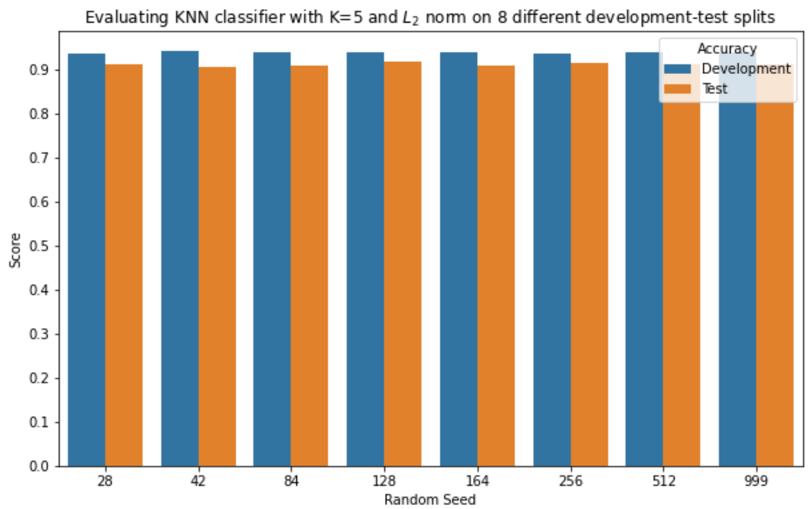


Average development accuracy score: 0.961

Average test accuracy score: 0.932

As we can see, MLE classifier performed well on the data it was trained on but also generalized well on unseen data.

(ii) Using a similar approach as above, we evaluated the KNN classifier and obtained the following results:

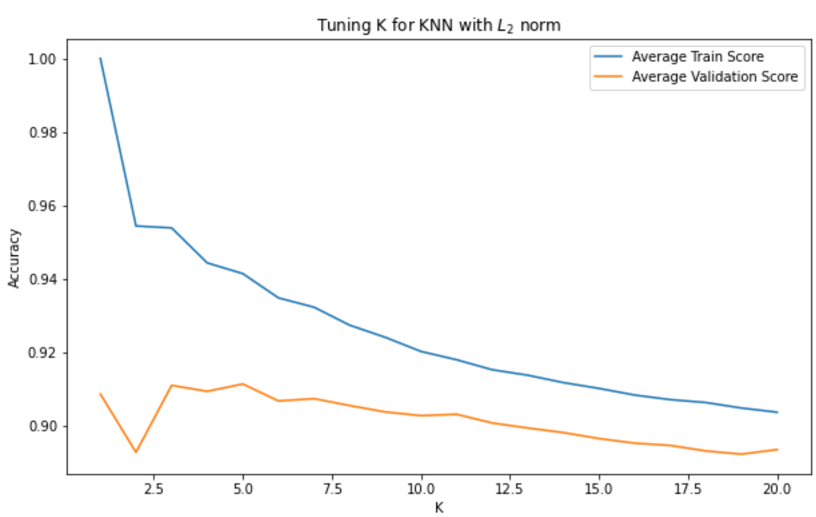


Average development accuracy score: 0.941

Average test accuracy score: 0.913

As we can see, KNN classifier with $K = 5$ and metric set to L_2 norm performed well on the data it was trained on and also generalized well on unseen data.

With KNN metric set to L_2 norm, we tuned the hyper-parameter K using the approach discussed above and obtained the following results:



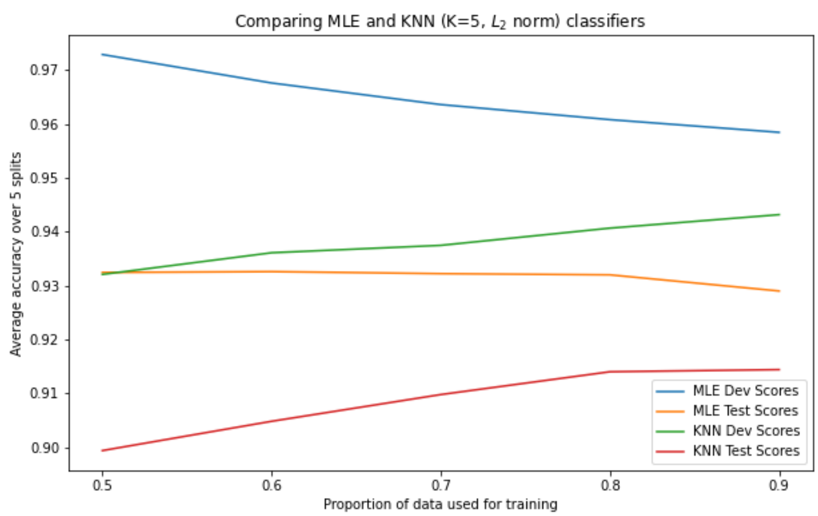
We can see that $K = 5$ seems to be the sweet spot. Training with $K = 5$ on the whole of development set, gave the following result:

Development accuracy score: 0.943

Test accuracy score: 0.908

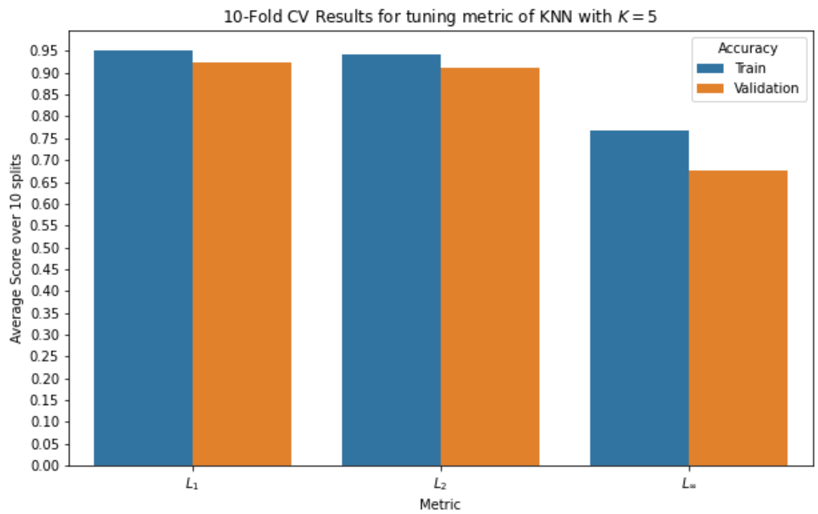
We will use $K = 5$ to compare KNN classifier with MLE classifier

(iii) We compared the MLE classifier with KNN classifier ($K = 5$ and L_2 norm) for varying training data size.



We can see that MLE classifier always performs better than KNN classifier. Even with 50% of training data, both the classifiers manage to do extremely well on unseen data.

(iv) We again follow the same approach of tuning hyper-parameter. This time our hyper-parameter is the metric to be used in KNN classifier. With $K = 5$, we tune the metric and obtain the following results:



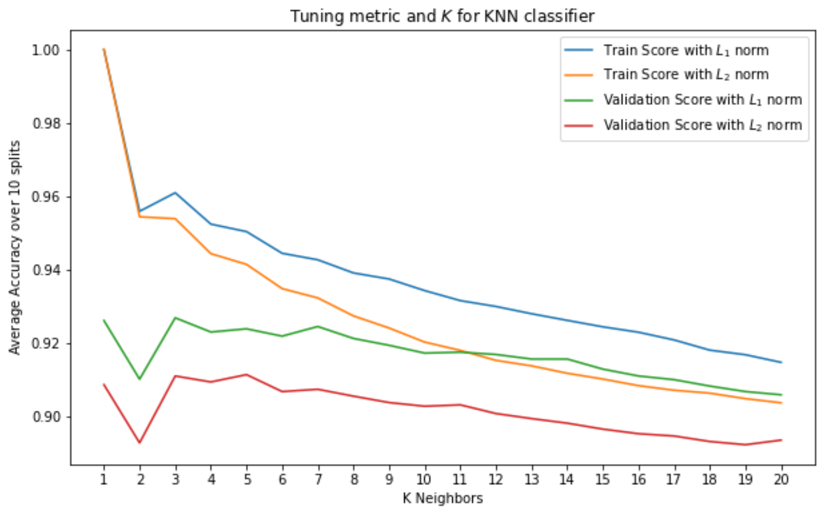
L_1 - Avg Train Score:0.950, Avg Validation Score:0.924
L_2 - Avg Train Score:0.941, Avg Validation Score:0.911
L_∞ - Avg Train Score:0.768, Avg Validation Score:0.675

We can see that KNN with L_∞ norm performs significantly worse than with the other metrics. Both models with L_1 and L_2 norm are comparable, however, L_1 wins by a small margin. We evaluated their performance by training on the whole of development set and testing on the test set.

L_1 Test Accuracy: 0.922
L_2 Test Accuracy: 0.908
L_∞ Test Accuracy: 0.689

Hence, we can go with L_1 as our metric for KNN classifier.

We can also tune both the metric and K for KNN classifier. Because of results obtained above, we will not consider L_∞ in search space. We obtained the following results:



This demonstrates that L_1 norm outperforms L_2 for optimal values of K. The optimal value of K for L_2 again turns out to be $K = 5$, as we obtained before. The optimal value of K for L_1 norm turns out to be $K = 7$.

Lastly, we evaluated our models with $(K = 5,\ L_2)$ and $(K = 7,\ L_1)$ on test dataset. We obtained the following results:

L_2 with $K = 5$ Test Accuracy: 0.908
L_1 with $K = 7$ Test Accuracy: 0.9235

Hence, L_1 norm with $K = 7$ is the best combination for the KNN classifier.