CSCI 567 Fall 2016
Homework 1

MohammedJunaid Hundekar
2548-4069-96
hundekar@usc.edu
Collaborators: Sagar Makhwana, Dhawal Shah, Ankit Kothari

September 21, 2016

1 Density Estimation

a. By definition of Beta distribution

$$f(x) = \frac{x^{\alpha-1}(1-x)^{(\beta-1)}}{B(\alpha,\beta)} \tag{1}$$

$$B(\alpha,\beta) = \frac{\tau(\alpha)\tau(\beta)}{\tau(\alpha+\beta)} \tag{2}$$

$$\tau(\alpha) = \tau(\alpha-1)! \tag{3}$$

$$\text{Given } \beta = 1; \alpha = ? \tag{4}$$

$$B(\alpha,1) = \frac{\tau(\alpha)\tau(1)}{\tau(\alpha+1)} \tag{5}$$

$$= \frac{(\alpha-1)!}{\alpha!} \tag{6}$$

$$= \frac{1}{\alpha} \tag{7}$$

using 7 to simplify 1

$$f(x) = \alpha x^{\alpha-1} \tag{8}$$

Taking log likelihood

$$l(\alpha) = \sum_{i=1}^{N} \log(\alpha x^{\alpha-1}) \tag{9}$$

$$= N \log \alpha + \sum_{i=1}^{N} (\alpha-1) \log x \tag{10}$$

Differentiating with respect to α and then equate to 0 to get Max

$$\frac{dl(\alpha)}{d\alpha} = \frac{N}{\alpha} + \sum_{i=1}^{N} \log(x) \tag{11}$$

$$\frac{dl(\alpha)}{d\alpha} = 0 \tag{12}$$

$$\frac{N}{\alpha} + \sum_{i=1}^{N} \log(x) = 0 \tag{13}$$

$$\alpha = -\frac{N}{\sum_{i=1}^{N} \log(x)} \tag{14}$$

Assume samples are generated from a normal distribution $N(\theta, \theta)$

$$f(x) = \frac{1}{\sigma\sqrt[2]{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{15}$$

given $\mu = \theta$ and $\sigma^2 = \theta$

$$f(x) = \frac{1}{\sqrt[2]{2\pi\theta}} e^{-\frac{(x-\theta)^2}{2\theta}} \tag{16}$$

Taking log likelihood

$$l(\theta) = \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt[2]{2\pi\theta}} e^{-\frac{(x-\theta)^2}{2\theta}}\right) \tag{17}$$

$$= -N\log(\sqrt[2]{2\pi\theta}) - \sum_{i=1}^{N} \frac{(x-\theta)^2}{2\theta} \tag{18}$$

$$= \frac{-1}{2}\left[\frac{N}{\theta} + N - \frac{\sum_1^N x^2}{\theta^2}\right] \tag{19}$$

For MLE set $\frac{dl(\theta)}{d\theta} = 0$

$$\frac{N}{\theta} + N - \frac{\sum_1^N x^2}{\theta^2} = 0 \tag{20}$$

$$N\theta^2 + N\theta - \sum_{1}^{N} x^2 = 0 \tag{21}$$

$$\theta = \frac{-N \pm \sqrt[2]{N^2 + 4N\sum_1^N x^2}}{2N} \tag{22}$$

$$\theta = \frac{1}{2}\left[-1 \pm \sqrt[2]{1 + \frac{4\sum_1^N x^2}{N}}\right] \tag{23}$$

b. Given KDE

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{k}K\left(\frac{x - X_i}{h}\right) \tag{24}$$

By linearity of expectation

$$E[\hat{f}(x)] = \frac{1}{Nh}\sum_{i=1}^{N}E\left[K\left(\frac{x - X_i}{h}\right)\right] \tag{25}$$

$$= N\frac{1}{N}\int_{-\infty}^{\infty}\frac{1}{h}K\left(\frac{x - X_i}{h}\right)f(X_i)dX_i \tag{26}$$

Using the substitution $X_i = t$ and $z = \frac{x-t}{h}$ in 26 and simplifying

$$E[\hat{f}(x)] = \int_{-\infty}^{\infty}K(z)f(x - hz)dz \tag{27}$$

Using Taylor's theorem on 27

$$E[\hat{f}(x)] = \int_{-\infty}^{\infty}K(z)dz\left[f(x) - hzf'(x) + \frac{h^2z^2}{2!}f''(x) - \frac{h^3z^3}{3!}f'''(x) + ...\right] \tag{28}$$

Using the properties of kernel function to simplify 27

$$E[\hat{f}(x)] = f(x) + \frac{h^2}{2}f''(x)\int_{-\infty}^{\infty}z^2K(z)dz + \frac{h^3z^3}{3!}f'''(x) + O(h^2) \tag{29}$$

$$E[\hat{f}(x)] - f(x) = \frac{h^2\sigma_k^2f''(x)}{2} + O(h^2) \tag{30}$$

2 Naive Bayes

a. Given the following
$P(Y=1)=\pi$
$P(X_j|Y=y_k)$ follows a Gaussian distribution $N(\mu_{ij},\sigma_j^2)$
By conditional independence the equation for $P(Y=1|X)$ simplifies to below from

$$P(Y=1|x)=\frac{\pi\prod P(X=x_i|Y=1)}{\pi\prod P(X=x_i|Y=1)+(1-\pi)\prod P(X=x_i|Y=0)} \tag{31}$$

$$=\frac{1}{1+\frac{1-\pi}{\pi}\frac{\prod P(X=x_i|Y=0)}{\prod P(X=x_i|Y=1)}} \tag{32}$$

$$\frac{P(X=x_i|Y=0)}{P(X=x_i|Y=1)}=\left[\frac{\frac{1}{\sigma_i\sqrt[2]{2\pi}}e^{-\frac{(X_i-\mu_{i0})^2}{2\sigma_j^2}}}{\frac{1}{\sigma_i\sqrt[2]{2\pi}}e^{-\frac{(X_i-\mu_{i1})^2}{2\sigma_j^2}}}\right] \tag{33}$$

$$=e^{\left[\frac{(X_i-\mu_{i1})^2-(X_i-\mu_{i0})^2}{2\sigma_i^2}\right]} \tag{34}$$

$$=e^{\left[X_i\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}+\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right]} \tag{35}$$

Using 35 to simplify 32

$$P(Y=1|x)=\frac{1}{1+exp\left[\ln\frac{1-\pi}{\pi}+\sum_i\left(X_i\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}+\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}\right)\right]} \tag{36}$$

$$P(Y=1|x)=\frac{1}{1+exp(\omega_o+\omega^TX)} \tag{37}$$

Comparing we 36 with 37 we get

$$\omega_o=\ln(\frac{1-\pi}{\pi})+\sum_i(\frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}) \tag{38}$$

$$\omega^T=\sum_i(\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}) \tag{39}$$

b. The joint probability distribution of Naive Bayes is given by

$P(X = x, Y = y_k) = P(Y = y_k) \prod_{j=1}^{D} P(X_j = x_j | Y = y_k)$

From the assumptions of 2.1; Taking the log likelihood

$$l(\theta) = \sum_{i=1}^{N} P(Y = y_n) + \sum_{i=1}^{N} \sum_{j=1}^{D} P(X = x_{ij} | Y = y_j) \tag{40}$$

We can now maximize/estimate the two terms separately

$$\sum_{i=1}^{N} P(Y = y_n) = \sum_{i=1}^{C} P(Y = y_n) N_c \tag{41}$$

$$\pi = \frac{N_{C=1}}{N} \tag{42}$$

$$1 - \pi = \frac{N_{C=0}}{N} \tag{43}$$

$$\sum_{i=1}^{N} \sum_{j=1}^{D} P(X = x_{ij} | Y = y_j) = \sum_{i=1}^{N} \sum_{j=1}^{D} \left(\frac{-\log(\sigma_{jY_c}^2 2\pi)}{2} - \frac{(X_{ji} - \mu_{jY_c})^2}{2\sigma_{jY_c}^2} \right) \tag{44}$$

Taking Partial derivative wrt μ_{jY_c} and simplifying:

$$\mu_{jY_c} = \sum_{i=1}^{N} \sum_{j=1}^{D} \frac{X_{ij}}{N} \tag{45}$$

Taking Partial derivative wrt $\sigma_{jY_c}^2$ and simplifying:

$$\sigma_{jY_c}^2 = \sum_{i=1}^{N} \sum_{j=1}^{D} \frac{(X_{ij} - \mu_{jY_c})^2}{N} \tag{46}$$

3 Nearest Neighbor

a. Given:
Unknown Point (20,7)

$\bar{x} = \frac{1}{N}\sum_i^n x_i = 12.769$ $\quad$ $\bar{y} = \frac{1}{N}\sum_i^n y_i = 12.307$

$\sigma_x = \sqrt[2]{\frac{1}{N-1}\sum_i^n (x_i - \bar{x})^2} = 20.716$ $\quad$ $\sigma_y = \sqrt[2]{\frac{1}{N-1}\sum_i^n (y_i - \bar{z})^2} = 24.913$

After Normalizing the data and calculating the L1 and L2 dist we get:

Table 1: Normalized points; L1 and L2 distance from (20,7)

Subject	**x1**	**y1**	**xn**	**yn**	**l2 dist**	**l1 dist**
Maths	0	49	-0.616	1.473	1.943	2.651
	-7	32	-0.954	0.790	1.645	2.307
	-9	47	-1.051	1.393	2.130	3.005
EE	29	12	0.783	-0.012	0.479	0.635
	49	31	1.749	0.750	1.699	2.363
	37	38	1.170	1.031	1.491	2.065
CS	8	9	-0.230	-0.133	0.585	0.660
	13	-1	0.011	-0.534	0.466	0.659
	-6	-3	-0.906	-0.614	1.318	1.656
	-21	12	-1.630	-0.012	1.989	2.180
Eco	27	-32	0.687	-1.778	1.601	1.903
	19	-14	0.301	-1.056	0.844	0.891
	27	-20	0.687	-1.297	1.135	1.422

Table 2: Nearest Neighbor Based on K

K	**L1**	**L2**
1	EE	CS
5	Tie(Eco,CS) Choose CS	Tie(Eco,CS) Choose CS

Since there is a tie for K=5 for both L1 and L2 distances between Economics and Computer Science, we choose the label of point which which is nearest to our query point to break the tie.

b. Given Volume = V with K points
N Points inside entire space, N_c points of each class C. With k_c points of class C inside sphere

$N = \sum_c N_c$

$K = \sum_c K_c$

$P(x|Y = c) = \frac{K_c}{N_c V}$

$$P(x) = \sum_c P(x|Y = c)P(Y = c) \tag{47}$$

$$= \frac{K_1}{N_1 V}\frac{N_1}{N} + \frac{K_2}{N_2 V}\frac{N_2}{N} \ldots + \frac{K_c}{N_c V}\frac{N_c}{N} \tag{48}$$

$$= \frac{1}{NV}\sum_c K_c \tag{49}$$

$$P(x) = \frac{K}{NV} \tag{50}$$

Using Bayes Rule:

$$P(Y = c|x) = \frac{P(x|Y = c)P(Y = c)}{P(x)} \tag{51}$$

$$= \frac{\frac{K_c}{N_c V}\frac{N_c}{N}}{\frac{K}{NV}} P(Y = c|x) \qquad = \frac{K_c}{K} \tag{52}$$

4 Decision Tree

a. By definition of entropy and conditional entropy

$H(X) = -\sum_{i=1}^{m} p_i \log p_i$

$H(Y|X) = -\sum_{i=1}^{m} p(X = x_i) H(Y|X = x_i)$

$H(Y|X) = -\sum_{mn} p(x_i, y_i) \log p(y_i|x_i)$

$I(Y; X) = H(Y) - H(Y|X)$

$$H(Acc) = -\left[p_{high} \log p_{high} + p_{low} \log p_{low}\right] \tag{53}$$
$$= -\left[\frac{73}{100} \log \frac{73}{100} + \frac{27}{100} \log \frac{27}{100}\right] \tag{54}$$
$$= 0.8414 \tag{55}$$
$$H(Acc|Weather) = -\left[\frac{28}{100}\left[\frac{28}{100} \log \frac{23}{28} + \frac{5}{28} \log \frac{5}{28}\right] + \frac{72}{100}\left[\frac{50}{72} \log \frac{50}{72} + \frac{22}{72} \log \frac{22}{72}\right]\right] \tag{56}$$
$$= 0.8288 \tag{57}$$
$$I(Acc; Weather) = 0.0125 \tag{58}$$

$$H(Acc|Traffic) = -\left[\frac{27}{100}\left[\frac{27}{27} \log \frac{27}{27}\right] + \frac{73}{100}\left[\frac{73}{73} \log \frac{73}{73}\right]\right] \tag{59}$$
$$= 0 \tag{60}$$
$$I(Acc; Traffic) = 0.8414 \tag{61}$$

By comparing -60 with -57 we can see that we get the highest gain when we split on **Traffic.**

b. Since we are normalizing is a linear function this does not change where the split occurs. The relative ordering of the tree will remain the same. Applying linear transformation will not affect the information to be gained from a split.

c. Given the definitions of Gini Index and Cross Entropy.
We need to prove that:

$\sum_{k=1}^{k} p_k(1 - p_k) \leq -\sum_{k=1}^{k} p_k \log p_k$

It is sufficient to prove the above inequality holds true for each term in the summation

$$p_k(1 - p_k) \leq p_k \log p_k \tag{62}$$

The above function ranges between [0,1]

Similarly, consider the function $p(x) = 1 - x + \log(x)$ defined in the same range [0,1].

To maximize $p(x)$

$p'(x) = 1 - \frac{1}{x} = 0$
$x = 1$

Thus the $p(x) \leq 0$

$1 - x + \log(x) \leq 0$

$1 - x + \leq -\log(x)$

$x(1 - x) + \leq -x\log(x)$

We see that the above equation is equal to 62.

This we can conclude that Gini Index $\leq$ Cross Entropy

5 Programming

Looking at some sample data and what each of the columns represents.

Table 3: Sample Data

1	**1.52101**	**13.64**	**4.49**	**1.10**	**71.78**	**0.06**	**8.75**	**0.00**	**0.00**	**1**
2	**1.51761**	**13.89**	**3.60**	**1.36**	**72.73**	**0.48**	**7.83**	**0.00**	**0.00**	**1**
3	**1.51618**	**13.53**	**3.55**	**1.54**	**72.99**	**0.39**	**7.78**	**0.00**	**0.00**	**1**
4	**1.51766**	**13.21**	**3.69**	**1.29**	**72.61**	**0.57**	**8.22**	**0.00**	**0.00**	**1**
5	**1.51742**	**13.27**	**3.62**	**1.24**	**73.08**	**0.55**	**8.07**	**0.00**	**0.00**	**1**
6	**1.51596**	**12.79**	**3.61**	**1.62**	**72.97**	**0.64**	**8.07**	**0.00**	**0.26**	**1**
7	**1.51743**	**13.30**	**3.60**	**1.14**	**73.09**	**0.58**	**8.17**	**0.00**	**0.00**	**1**
8	**1.51756**	**13.15**	**3.61**	**1.05**	**73.24**	**0.57**	**8.24**	**0.00**	**0.00**	**1**
9	**1.51918**	**14.04**	**3.58**	**1.37**	**72.08**	**0.56**	**8.30**	**0.00**	**0.00**	**1**

Table 4: Column Attributes

1	Id number: 1 to 214
2	RI: refractive index
3	Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)
4	Mg: Magnesium
5	Al: Aluminum
6	Si: Silicon
7	K: Potassium
8	Ca: Calcium
9	Ba: Barium
10	Fe: Iron
11	Type of glass: (class attribute)

We can see that since column 1 is Id Number and 11 is class label, we are effectively left with 9 attributes.

Table 5: Attributes and their correlation to class

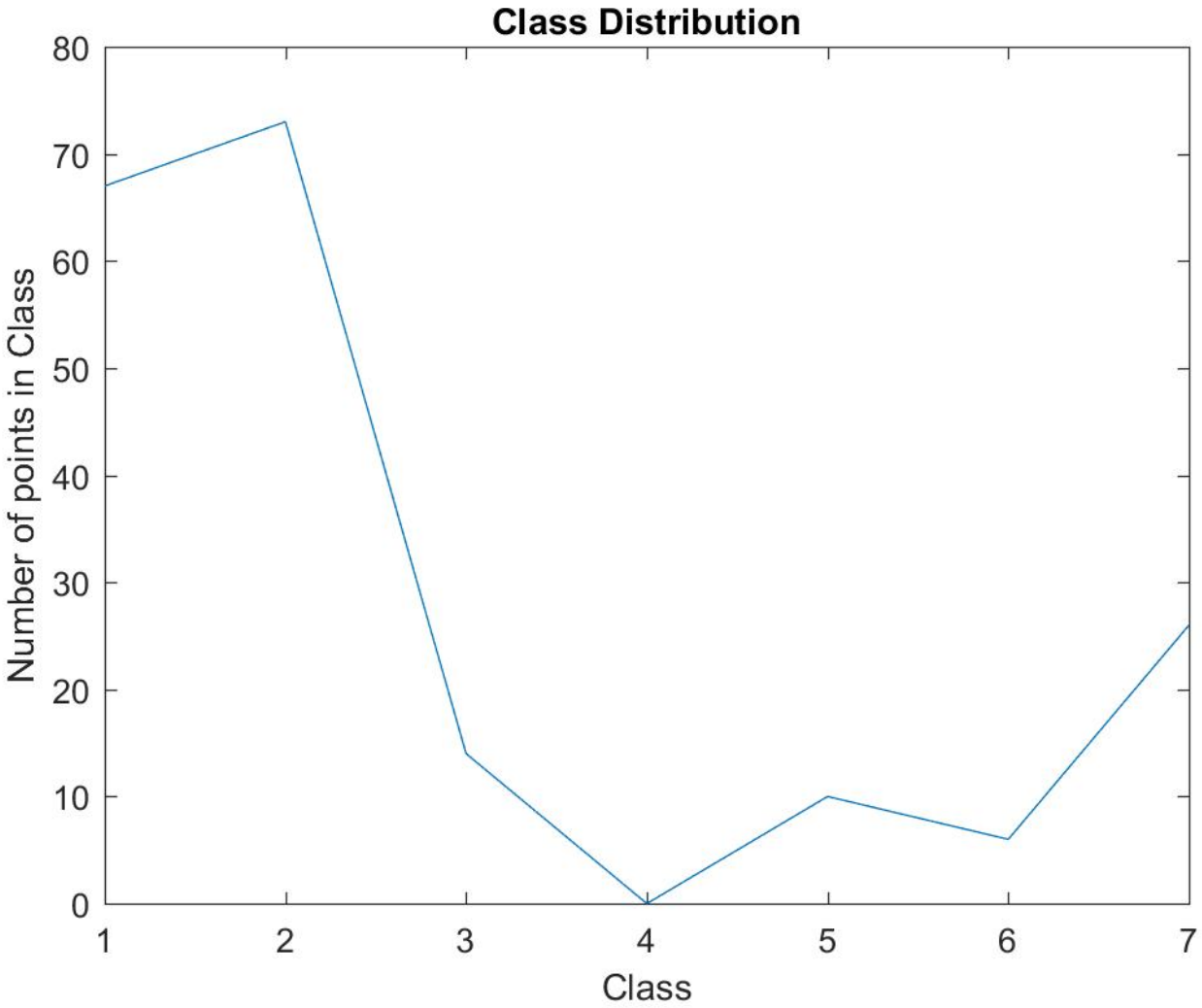
Attribute:	**Name**	**Min**	**Max**	**Mean**	**SD**	**Correlation with class**
2	RI:	1.5112	1.5339	1.5184	0.003	-0.1642
3	Na:	10.73	17.38	13.4079	0.8166	0.503
4	Mg:	0	4.49	2.6845	1.4424	-0.7447
5	Al:	0.29	3.5	1.4449	0.4993	0.5988
6	Si:	69.81	75.41	72.6509	0.7745	0.1515
7	K:	0	6.21	0.4971	0.6522	-0.01
8	Ca:	5.43	16.19	8.957	1.4232	0.0007
9	Ba:	0	3.15	0.175	0.4972	0.5751
10	Fe:	0	0.51	0.057	0.0974	-0.1879

In the above table we can see that attribute 8 Ca has very low correlation to the class label so we can say that it will not affect the categorization of any point

Table 6: Distribution of class

Class Label	**Count**	**Percentage**
1	67	34.184
2	73	37.245
3	14	7.143
5	10	5.102
6	6	3.061
7	26	13.265

The above table shows how the distribution of the glass types in the training set. Class 2 with 72 rows is the majority class followed closely by Class 1 with 67 rows.



Looking at the plot we can clearly see that this is not a uniform distribution.

a. kNN Performance

The performance of kNN on training set using Leave One Out strategy is given in the table below

Table 7: kNN Performance on Training Set

K	**L1/L2**	**Count**	**Accuracy**
1	L1	147	75
1	L2	139	70.91836735
3	L1	145	73.97959184
3	L2	141	71.93877551
5	L1	134	68.36734694
5	L2	131	66.83673469
7	L1	135	68.87755102
7	L2	132	67.34693878

The performance of kNN on testing data is given in the table below

Table 8: kNN Performance on Testing Set

K	**L1/L2**	**Count**	**Accuracy**
1	L1	12	66.66666667
1	L2	11	61.11111111
3	L1	11	61.11111111
3	L2	11	61.11111111
5	L1	10	55.55555556
5	L2	10	55.55555556
7	L1	9	50
7	L2	10	55.55555556

b. Naive Bayes Performance

The performance of Naive Bayes is given in the below Table

Table 9: Naive Bayes Performance

	Training	**Testing**
Accuracy	54.591	33.333

c. Performance Comparison

By comparing the above Table 6, 7 and 8. We can clearly observe that kNN performs much better compared Naive Bayes for both the training and testing data.
This is due to the fact that kNN makes less assumptions compared to Naive Bayes, i.e L1 and L2 nor do not make any assumptions and are strictly empirical in nature.
Whereas Naive Bayes assumes conditional independence; and along with this we have also made the assumption that the attributes follow a Gaussian distribution and generated the pdf from a mere 196 samples.
These assumptions are not usually true in the real world and may not be true for this case as well.
We get a better performance on the training data compared to testing, for both kNN and Naive Bayes due to over fitting of our model.