CSCI 567 Assignment 1
Fall 2016

Snehal Adsule
2080872073
adsule@usc.edu

September 21,2016

1 Problem 1

1.1 1 (a) 1

Given that $X_i \sim Beta(\alpha, \beta)$, and $\beta = 1$

$$\begin{aligned} f(x_i) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} && (1) \\ &= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\Gamma(1)} x^{\alpha-1} && \Gamma(\alpha+1) = \alpha\Gamma(\alpha) \quad (2) \\ &= \frac{\alpha\Gamma(\alpha)}{\Gamma(\alpha)} x^{\alpha-1} && \Gamma(\alpha) = 1 \quad (3) \\ &= \alpha x^{\alpha-1} && (4) \end{aligned}$$

We need to calculate Maximum likelihood MLE for α: Consider $X = (X_1, X_2, \ldots, X_n)$,likelihood function as $L(\alpha|X)$
$L(\alpha|X) = \prod_{i=1}^{n} f(x_i)$

$$=> L(\alpha|X) = \big(\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\Gamma(1)}\big)^n \prod_{i=1}^{n}(x_i)^{\alpha-1} \quad (5)$$

Log likelihood :

$$LL(\alpha|X) = \log(L(\alpha|X)) = n\log(\alpha) + (\alpha-1)\sum_{i=1}^{n} x_i \quad (6)$$

$$\frac{dLL}{d\alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n} \log(x_i) => 0 \quad (7)$$

$$n + \alpha\sum_{i=1}^{n} \log x_i = 0 \implies \hat{\alpha} = \frac{n}{\sum_{i=1}^{n} log(1/x_i)} \quad (8)$$

As log is a concave function, we have minima at $\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} log(1/x_i)}$

Therefore,

$$\hat{\alpha}_{LL} = \frac{n}{\sum_{i=1}^{n} log(1/x_i)} \quad (9)$$

1.2 1 (a) 2

Given: $x_i \sim N(\mu = \theta, \sigma^2 = \theta)$ =¿ $f(x_i) = \frac{1}{\sqrt(2\pi\theta)} e^{-\frac{(x_i-\theta)^2}{2\theta}}$
MLE estimate for θ:

$$MLE(\theta|X) = (\frac{1}{\sqrt(2\pi\theta)})^N \prod_{i=1}^{N} e^{-\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\theta}}$$

$$LL(\theta|X) = \log(L(\theta|X)) = -\frac{N}{2}\log((2\pi\theta)) - \sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\theta}$$

$$\frac{dLL}{d\theta} = -\frac{N}{2}(\frac{1}{\theta}) + \frac{\sum_{i=1}^{n} x_i^2}{2\theta^2} - \frac{N\theta}{2}$$

$$\frac{dLL}{d\theta} = 0 \implies N\theta^2 + N\theta - \sum_{i=1}^{n} x_i^2 = 0$$

Solving quadratic equation for θ and we have two solutions,but $\theta = \sigma^2 => \theta \geq 0$

$$\theta = \frac{-N \pm \sqrt{N^2 + 4N\sum_{i=1}^{n} x_i^2}}{2N}$$

$\implies \hat{\theta_{MLE}} = \frac{-N+\sqrt{N^2+4N\sum_{i=1}^{n} x_i^2}}{2N}$ (as $\hat{\theta}$ is positive).

1.3 1 (b) 1

Given KDE in the form of : $\hat{f(x)} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x-X_i}{h})$

$$\begin{aligned}
E[\hat{f(x)}] &= E[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x-X_i}{h})] \\
&= \frac{1}{nh}E[K(\frac{x-X_i}{h})] && E[f(X1)] = E[f(X2)] \\
&= \frac{1}{h}E[K(\frac{x-X_1}{h})] \\
&= \frac{1}{h}E[K(\frac{x-t}{h})] \\
&= \frac{1}{h}\int K(\frac{x-t}{h})f(t)dt \qquad (byE[g(x)] = \int g(x)f(x)dxdefinition)
\end{aligned}$$

1.4 1 (b) 2

Now, $z = \frac{x-t}{h} \implies t = x - zh$ Therefore,

$$E[f\hat{(x)}] = \frac{1}{h}\int K(z)f(x-hz)dz$$

using Taylors Theorem:

$$f(x-hz) = f(x) - f'(x)hz + \frac{1}{2}f''(x)\frac{(hz)^2}{2} - \frac{1}{3}f'''(x)\frac{(hz)^3}{3!} + \dots$$
$$+ (-1)^n\frac{1}{n!}f^{(n)}(x)\frac{(-hz)^n}{n!} + O(h^{n+1})$$

By definition, $\int K(z)dz = 1$ and $\int zK(z)dz = 0$.

$$\int K(z)f(x-hz)dz = f(x) - hf'(x)\int zK(z)dz + \frac{1}{2}(h^2)f^{''}(x)\int z^2K(z)dz + \cdots + (-1)^n\frac{1}{n!}f^{(n)}(x)\int z^nK$$

Now,

$$Bias = E[f\hat{(x)}] - f(x) = \frac{1}{2}(h^2)f^{''}(x)\int z^2K(z)dz + \cdots + O(h^{n+1}), using \int K(z)f(x-hz)dz$$

Therefore, as $h \longrightarrow 0$, $Bias \longrightarrow 0$

$$Bias = \frac{1}{2}(h^2)f^{''}(x)\int z^2K(z)dz + \cdots + O(h^{n+1})$$

2 . Naive Bayes :

2.1 2. (a) Naive Bayes in Logistic Regression form

Gaussian Naive Bayes as the joint distribution:

$$\begin{aligned} P(X=x, Y=y) &= P(Y=y)P(X=x|Y=y) \\ &= P(Y=y)\prod_{d=1}^{D}P(X_d=x_d|Y=y) \end{aligned}$$

Given that $P(Y=1)=\pi => P(Y=0)=1-\pi$ .....(1)
$P(X_j|Y=y_k) \sim N(\mu_{jk},\sigma_{jk}^2)$ ......(2)

$$
\begin{aligned}
P(Y=1|X) &= \frac{P(X|Y=1)P(Y=1)}{P(X)} \\
&= \frac{P(X|Y=1)P(Y=1)}{P(X|Y=1)P(Y=1)+P(X|Y=0)P(Y=0)} \\
&= \frac{1}{1+\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}} \\
&= \frac{1}{1+\exp(\log(\frac{P(X|Y=0)P(Y=0)}{P(X|Y=1)P(Y=1)}))} \\
&= \frac{1}{1+\exp(\log(P(X|Y=0)P(Y=0))-\log(P(X|Y=1)P(Y=1)))} \\
&= \frac{1}{1+\exp(-(\log(\frac{P(Y=1)}{P(Y=0)}))+\log(P(X|Y=0))-\log(P(X|Y=1)))} && using(1) \\
&= \frac{1}{1+\exp(-(\log(\frac{\pi}{1-\pi}))+\log(P(X|Y=0))-\log(P(X|Y=1)))} && ....(3)
\end{aligned}
$$

Using fact (2),

$$
\begin{aligned}
\log(P(X|Y=0)) &= -\frac{\log(2\pi\sigma_j^2)}{2}-\frac{(x_j-\mu_{j0})^2}{2\sigma_j^2} \\
\log(P(X|Y=1)) &= -\frac{\log(2\pi\sigma_j^2)}{2}-\frac{(x_j-\mu_{j1})^2}{2\sigma_j^2}
\end{aligned}
$$

Subtracting we get

$$
\begin{aligned}
\log(P(X|Y=0))-\log(P(X|Y=1)) &= -\log(\frac{2\pi\sigma_j^2}{2\pi\sigma_j^2})-\frac{(x_j-\mu_{j0})^2}{2\sigma_j^2}+\frac{(x_j-\mu_{j1})^2}{2\sigma_j^2} \\
&= -\frac{(x_j^2-2x_j\mu_{j0}^2+\mu_{j0}^2)}{2\sigma_j^2}+\frac{(x_j^2-2x_j\mu_{j1}^2+\mu_{j1}^2)}{2\sigma_j^2} \\
&= -\frac{(\mu_{j0}^2)-\mu_{j1}^2}{2\sigma_j^2}+\frac{(-2x_j\mu_{j0}^2+2x_j\mu_{j1}^2)}{2\sigma_j^2}
\end{aligned}
$$

Substituting in (3)

$$
P(Y=1|X) = \frac{1}{1+\exp(-(\log(\frac{\pi}{1-\pi}))-\frac{(\mu_{j0}^2)-\mu_{j1}^2}{2\sigma_j^2}+\frac{(-\mu_{j0}^2+\mu_{j1}^2)x_j}{\sigma_j^2}))}
$$

We can compare this with the given form and calulcate w_0

$$w_0 = \log(\frac{\pi}{1-\pi})) + \sum_{j=1}^{D} \frac{(\mu_{j0}^2 - \mu_{j1}^2)}{2\sigma_j^2}$$
$$w = \sum_{j=1}^{D} \frac{(-\mu_{j0}^2 + \mu_{j1}^2)}{\sigma_j^2}$$

2.2 2(b) MLE for GNB

Consider the likelihood function

$$\begin{aligned} L(\mu, \sigma, p|(X,Y)) &= \prod_{1}^{N} P(Y_i = y_i) \times \prod_{j=1}^{D} P(X_i = x_{ij}|Y = y_i) \\ \log(L) &= \sum_{i=1}^{N} (\log(P(Y_i = y_i)) + \sum_{i=1}^{N}\sum_{j=1}^{D} \log(P(X_i = x_{ij}|Y = y_i))) \\ &= \sum_{i=1}^{N} \log(P(Y_i = y_i)) + \sum_{i=1}^{N}\sum_{j=1}^{D} \log(P(X_{ij} = x_{ij}|Y = y_i)) \\ &= \sum_{k=1}^{K} log(P(Y = k)) \times N_k + \sum_{k=1}^{K}\sum_{j=1}^{D} \log(P(X_{ij} = x_{ij}|Y = k)) \times N_k \\ &= (\log(\pi) + \log(1-\pi)) \times 2 + \sum_{k=1}^{K}\sum_{j=1}^{D} \log(P(X_{ij} = x_{ij}|Y = k)) \times 2 \end{aligned}$$

Now,

$$\begin{aligned} \frac{\partial LL}{\partial p_k} &= 0 & \ldots (1) \\ \frac{\partial LL}{\partial \mu_{jk}} &= 0 & \ldots (2) \\ \frac{\partial LL}{\partial \sigma jk} &= 0 & \ldots (3) \end{aligned}$$

Solving equation 1 , $\frac{d(\log \pi + \log(1-\pi)}{d\pi} = 0 \implies \frac{1}{\pi} - \frac{1}{1-\pi} = 0$
$\implies \hat{\pi} = 0.5$

Solving equation 2,

$$\frac{\partial \sum_{k=1}^{K} \sum_{j=1}^{D} \log(P(X_i = x_j|Y = k)) \times N_k}{\partial \mu_{jk}} = 0$$

$$\frac{\sum_{j;Y_j=k}(x_{ij} - \mu_{jk})}{\sigma_{jk}} = 0 \implies (x_{j0} - \hat{\mu}_{jk}) + (x_{j1} - \hat{\mu}_{jk}) = 0$$

$$\hat{\mu_{jk}} = \frac{(x_{j0} + x_{j1})}{N_k} => \hat{\mu_{jk}} = \frac{\sum_{i;Y_i=k} x_{ij}}{N_k}$$

For equation 3,

$$\frac{\partial \sum_{k=1}^{K} \sum_{j=1}^{D} \log(P(X_i = x_{ij}|Y = k)) \times N_k}{\partial \sigma_{jk}} = 0$$

$$\frac{\partial}{\partial \sigma_{jk}^2} \sum_{i;Y_i=k} \Big(-\frac{\log(2\pi\sigma_{jk}^2)}{2} - \frac{(x_{ij} - x_{jk})^2}{2\sigma_{jk}^2} \Big) = 0$$

$$\sum_{i;Y_i=k} \Big(-\frac{1}{\sigma_{jk}} + \frac{(x_{ij} - \mu_{jk})^2}{2\sigma_{jk}^3} \Big) = 0 \implies \sum_{i;Y_i=k} (-\sigma_{jk}^2 + (x_{ij} - \mu_{jk})^2) = 0$$

$$\hat{\sigma_j} = \frac{\sum_{i;Y_i=k}(x_{ij} - \hat{\mu_{jk}})^2}{N_k}$$

3 Nearest Neighbour

3.1 3.(a) Student Major classification

Mean and Standard deviation are calculated for normalization
$\mu_x = 12.76923077, \mu_y = 12.30769231$
$\sigma_x = 20.71695701, \sigma_y = 25.93062737$
Classify point (20,7)
Normalized to (0.349026608,-0.204688156)

$L1 = |x_1 - x_0| + |y_1 - y_0|$
$L2 = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$

For k=1,
$L1 = 0.627$ classified as major = EE
$L2 = 0.457$ classified as major= CS

For k=5, Resolving ties by nearest majority
L1 EE1 0.627 <CS2 0.646 <CS1 0.656 <Eco2 0.858 <Eco3 1.379 $\implies$ CS
L2 CS2 0.457 <EE1 0.475 <CS1 0.584 <Eco2 0.811 <Eco3 1.094 $\implies$ CS

Intermediate Calculations are liste din the table

ID	x	y	$x-\mu_x$	$(x-\mu_x)^2$	$y-\mu_y$	$(y-\mu_y)^2$	xn	yn	L1	L2
M1	0	49	-12.77	163.05	36.69	1346.33	-0.62	1.42	2.59	1.89
M2	-7	32	-19.77	390.82	19.69	387.79	-0.95	0.76	2.27	1.62
M3	-9	47	-21.77	473.90	34.69	1203.56	-1.05	1.34	2.94	2.08
EE1	29	12	16.23	263.44	-0.31	0.09	0.78	-0.01	0.63	0.48
EE2	49	31	36.23	1312.67	18.69	349.40	1.75	0.72	2.33	1.68
EE3	37	38	24.23	587.13	25.69	660.09	1.17	0.99	2.02	1.45
CS1	8	9	-4.77	22.75	-3.31	10.94	-0.23	-0.13	0.66	0.58
CS2	13	-1	0.23	0.05	-13.31	177.09	0.01	-0.51	0.65	0.46
CS3	-6	-3	-18.77	352.28	-15.31	234.33	-0.91	-0.59	1.64	1.31
CS4	-21	12	-33.77	1140.36	-0.31	0.09	-1.63	-0.01	2.17	1.99
Eco1	27	-32	14.23	202.51	-44.31	1963.17	0.69	-1.71	1.84	1.54
Eco2	19	-14	6.23	38.82	-26.31	692.09	0.30	-1.01	0.86	0.81
Eco3	27	-20	14.23	202.51	-32.31	1043.79	0.69	-1.25	1.38	1.09

3.2 3(b)

Given that $\sum_c N_c = N$, where N is total points and N_c points with label class c.
Also, $p(x|Y=c) = \frac{K_c}{N_c V}$ and $\sum K_c = K$
Class prior as $p(Y=c) = \frac{N_c}{N}$

Unconditional density $p(x) = \sum_c p(x|Y=c)p(Y=c) = \sum_c \frac{K_c}{N_c V} \times \frac{N_c}{N}$
$= \sum_c \frac{K_c}{NV} = \frac{K}{NV}$
Posterior $P(Y=c|x) = \frac{P(x|Y=c) \times P(Y=c)}{P(x)}$
$= \frac{\frac{K_c}{N_c V} \times \frac{N_c}{N}}{\frac{K}{NV}} = \frac{K_c}{K}$

4 . Decision Tree

4.1 4 (a)Accident rate

It can be observed that the attribute Traffic, divides the Accident Rate into High and Low 100% for the given observations. However, weather doesn't give a clear idea. The decision tree are more intitutive to humans and this is one such example.This can be further supported by the caluclation on Information Gain
$IG = H[Y] - H[Y|X]$, where Y is the result variable and X is an attribute.

$$H[Y] = -p_{high}\log(p_{high}) - p_{low}\log(p_{low}) = -(0.73)log(0.73) - (.27)log(0.27) = 0.253_log_10 = 0.8407$$

$$
\begin{aligned}
H[Y = weather|X] =& - p_{sunny} \times p_{high,sunny} \log p_{high,sunny} \\
& + p_{sunny} \times p_{low,sunny} \log p_{low,sunny} \\
& - p_{rainy} \times p_{high,rainy} \log p_{high,rainy} \\
& + p_{rainy} \times p_{low,sunny} \log p_{low,sunny} \\
& = -\frac{28}{100} \times (\frac{23}{28} \log \frac{23}{28}) + \frac{28}{100} \times (\frac{5}{28} \log \frac{5}{28}) \\
& - \frac{77}{100} \times (\frac{55}{70} \log \frac{55}{70}) - \frac{77}{100} \times (\frac{22}{70} \log \frac{22}{70}) \\
& = 0.8983 \\
H[Y = traffic|X] =& - p_{heavy} \times p_{high,heavy} \log p_{high,heavy} \\
& + p_{heavy} \times p_{low,heavy} \log p_{low,heavy} \\
& - p_{light} \times p_{high,light} \log p_{high,light} \\
& + p_{light} \times p_{low,heavy} \log p_{low,heavy} \\
& = -\frac{73}{100} \times (\frac{73}{73} \log \frac{73}{73}) + \frac{73}{100} \times (\frac{0}{73} \log \frac{0}{73}) \\
& - \frac{27}{100} \times (\frac{0}{27} \log \frac{0}{27}) - \frac{27}{100} \times (\frac{27}{27} \log \frac{27}{27}) \\
& = 0
\end{aligned}
$$

As H[Y=traffic—X] <H[Y=weather—X] $\implies$ IG(traffic) >IG(weather) . Hence traffic is better candidate than weather, with maximum information gain=1.

4.2 4 (b)Different Decision Tree

The normalization will not change the entopy and information gain,and hence both $T1$ and $T2$ will be same, as long as the dataset is same.

4.3 4 (c)Gini and Cross Entropy

Gini index = $\sum_{k=1} K p_k(1 - p_k)$ and Cross Entropy = $-\sum_{k=1}^{K} p_k \log(p_k)$
Suppose we have a function $f(p_k)$ as
$f(p_k) = (1 - p_k) - (-\log p_k)$ such that $0 \leq p_k \leq 1$

Taking derivatives we can say,
$f'(p_k) = -1 + \frac{1}{p_k} = -\frac{1-p_k}{p_k} \leq 0 \forall p_k \in [0, 1]$

As, $f'(p_k)$ is a non-increasing function which $\implies$ $f(p_k) \geq f(1) \forall p_k \in (0, 1]$
$(1 - p_k) - (-\log p_k) \geq 0$ Multiply by p_k
$\implies p_k(1 - p_k) - (-p_k \log p_k) \geq 0$
$\implies p_k(1 - p_k) \geq -p_k \log p_k \implies$
$\sum_{k=1}^{K} p_k(1 - p_k) \geq \sum_{k=1}^{K} -p_k \log p_k$
Gini index is less than corresponding value of Cross Entropy
$\implies$ Gini index is better approximation of misclassification error

5 . Programming Glass Identification

5.1 4 (a)Data Inspection

There are 11 columns wiht 10 attributes and ID as indexing column for the data.
Yes, except ID being just indexing column
As per data decription on UCI archieve, there are 7 glass types but in the training and test there are only 6 classes. Type 4 data is not there in the dataset.

Class Distribution - Training Data

Glass type	Count	P(Glass)
1	67	0.341
2	73	0.372
3	14	0.071
5	10	0.051
6	6	0.030
7	26	0.132

We can observe that Class 1 and 2 are majority classes. Yes, it can be colnsidered as uniform distribution .Also, there are no missing points in the data.

5.2 4 (b)Naive Bayes

There are $\mu_{jk} = 0$ and σ^2_{jk}=0 e.g. Class 6, and calculating the conditional probability will be handled as 0 in these cases.

Naive Bayes Accuracy
Naive Bayes : Test Accuracy : 33.3333%
Naive Bayes : Train Accuracy: 54.5918%

5.3 5 (c) kNN Results

The data was normalized using $x_j = (x_j - \mu_j/\sigma_j)$
kNN Test Accuracy

k= 1	66.666	61.111
k= 3	61.111	61.111
k= 5	55.555	55.555
k= 7	50.0	50.0

kNN Training LOO Accuracy

k=1	75.0	71.4285714286
k=3	72.4489795918	70.4081632653
k=5	76.0204081633	72.4489795918
k=7	74.4897959184	72.4489795918

5.4 5 (d)Naive Bayes and kNN Comparison

Naive Bayes is generative and kNN is discriminative, this impacts majorly the performance of the classifiers.

It was observed that kNN performed better than Naive Bayes as it is not dependant on the generation of the data. The prior probablity in the Naive Bayes classifier is more bias towards the dominating classes while classifying. Therefore the correlation between aatribute may result in misleading results at times.

kNN is very slow compared to naive bayes classifier, though gave a better performance. For large dataset, it may take even more time.

kNN is not dependant on the conditional probabilty like Naive Bayes, which may sometimes lead to errors.

There are some points where $\mu = 0$ and $\sigma = 0$, this affects the naive bayes performace as well where as the normalization of kNN further boosts the performance.

Validating the Training accuracy with Leave One Out is better measure in kNN, which was missing with GNB classifier.