# Fall2016 HomeWork1

Hema Venkata Krishna Giri Narra

September 21, 2016

## 1 Density Estimation

### 1.1 MLE of Beta Distribution

Given that the i.i.d random variables are of Beta distribution with $\beta = 1$ we have:
Beta($\alpha$,$\beta$)= $\frac{x^{\alpha-1}1-x^{\beta-1}}{B(\alpha,\beta)}$
Since $\beta = 1$:
Beta($\alpha$,$\beta$)= $x^{\alpha-1}\alpha$
The log likelihood function for MLE of $\alpha$ for this PDF is:
$l(\alpha) = \sum_{i=1}^{N} \log(x_i^{\alpha-1}\alpha)$
Taking the derivative of $l$ w.r.t $\alpha$ and setting it to zero
$\frac{dl(\alpha)}{d\alpha} = \sum_{i=1}^{N} (\log(x_i) + \frac{1}{\alpha}) = 0$

$$\hat{\alpha} = \frac{-N}{\sum_{i=1}^{N} \log(x_i)}$$

### 1.2 MLE of Normal Distribution

$N(\theta, \theta) = \frac{1}{\sqrt{2\pi\theta}} \exp^{\frac{-(x_i-\theta)^2}{2\theta}}$
The log likelihood function for MLE of $\theta$ for this PDF is:
$l(\theta) = \sum_{i=1}^{N} (-\log\sqrt{2\pi\theta} - \frac{(x_i-\theta)^2}{2\theta})$
Taking the derivative of $l$ w.r.t $\theta$ and setting it to zero
$\frac{dl(\theta)}{d\theta} = 0$

$$\rightarrow N\theta^2 + N\theta - \sum_{i=1}^{N} x_i^2 = 0$$

$$\rightarrow \hat{\theta} = -\frac{1}{2} \pm \sqrt{N^2 + 4N(\sum_{i=1}^{N} {x_i}^2)}$$

To maximize $l(\theta)$, $\frac{d^2l}{d\theta} < 0$

$\rightarrow \theta < -\frac{1}{2}$

$\rightarrow \hat{\theta} = -\frac{1}{2} - \sqrt{N^2 + 4N(\sum_{i=1}^{N} {x_i}^2)}$

### 1.3 Kernel Density Estimation

By linearity of Expectation: $E[\hat{f}(x)] = \frac{1}{n}\sum_{i=1}^{n} E(\frac{1}{h}K(\frac{x-X_i}{h}))$
Since $X_i$ are independent and identically distributed according f(x)
$\sum_{i=1}^{n} E(\frac{1}{h}K(\frac{x-X_i}{h})) = nE(\frac{1}{h}K(\frac{x-X}{h}))$
$E[\hat{f}(x)]$
$= E(\frac{1}{h}K(\frac{x-X}{h})$
$= \frac{1}{h}\int K(\frac{x-t}{h})f(t)dt$
$= \int K(z)f(x-hz)dz$
By Taylor's theorem at x:
$= \int dzK(z)(f(x) + f'(x)(-hz)) + \frac{f''(x)}{2!}(-zh)^2 + \frac{f'''(x)}{3!}(-zh)^3..$
Applying properties of kernel function $\int zK(z)dz = 0, \int K(z)dz = 1$
$= f(x) + \frac{f''(x)}{2!}\int dzK(z)(-zh)^2 + O(h)^3..$
$\rightarrow Bias(\hat{f}(x) = E(\hat{f}(x) - f(x)) = \frac{f''(x)}{2!}h^2\int z^2K(z)dz + O(h)^3$
If $h \rightarrow 0$ then $Bias \rightarrow 0$

## 2 Naive Bayes Parameters estimation

### 2.1 a

$P(Y = 1/X) = \frac{P(\frac{X}{Y=1})P(Y=1)}{P(\frac{X}{Y=1})P(Y=1)+P(\frac{X}{Y=0})P(Y=0)}$
$\rightarrow \frac{1}{1+\frac{P(\frac{X}{Y=0})P(Y=0)}{P(\frac{X}{Y=1})P(Y=1)}}$
$\rightarrow \frac{1}{1+F}$
On Taking log and simplifying the term F
$\log T = \log\frac{1-\pi}{\pi} + \sum_{j=1}^{D}\frac{u_{j1}^2-u_{j0}^2}{2\sigma_j^2} + \sum_{j=1}^{D} x_j\frac{u_{j0}-u_{j1}}{\sigma_j^2}$
Taking $exp$ on both sides and then substituting F in the expression of $P(Y = 1/X)$ above
$\rightarrow w_0 = -(\log\frac{1-\pi}{\pi} + \sum_{j=1}^{D}\frac{u_{j1}^2-u_{j0}^2}{2\sigma_j^2})$
Not sure if **w** is expected in vector form. In this case
$\mathbf{w}^T = [\frac{u_{10}-u_{11}}{\sigma_1^2}\frac{u_{20}-u_{21}}{\sigma_2^2}....\frac{u_{D0}-u_{D1}}{\sigma_D^2}]$

### 2.2 b

To Maximize likelihood for the parameters take partial derivative of log likelihood w.r.t each and equate to 0.

$\hat{u_{jk}} = \frac{1}{D} \sum_{j=1}^{D} x_j$

$\hat{\sigma_{jk}} = \frac{1}{D} \sum_{j=1}^{D} (x_j - u_{jk})^2$

## 3 kNN

### 3.1 a

Mean of inputs is (12.769, 12.307)
Std. Deviation is (20.717, 25.93)
Normalized coordinates for (20,7) are (0.349, -0.204)
L2 metric, K = 1:
**Computer Science**, distance is 0.457
L2 metric , K = 5:
5 nearest points in order are Computer Science, Electrical Engineering, Computer Science, Economics, Economics
Tie and computer science is closest → **Computer Science**
L1 metric , K = 1:
**Electrical Engineering**, distance is 0.6272
L1 metric , K = 5:
5 nearest points in order are Electrical Engineering, Computer Science, Computer Science, Economics, Economics
Tie and closest data point belongs to Computer science → **Computer science**
For K=5 both metrics give the same major
For K=1 both metrics give different majors
L1 metric gives different majors when K=1, K=5
L2 metric gives the same major in both K=1,5

### 3.2 b

Unconditional density using total probability theorem

$p(x) = \sum_c p(\frac{x}{Y=c}) P(Y = c)$

$= \sum_c \frac{K_c}{N_c V} \frac{N_c}{N}$

$= \frac{K_c}{VN}$

$= \frac{K}{VN}$

Posterior probability by Bayes rule:

$p(\frac{Y=c}{x}) = \frac{p(\frac{x}{Y=c}) p(Y=c)}{p(x)}$

$= \frac{\frac{K_c}{NV}}{\frac{K}{VN}}$
$= \frac{K_c}{K}$

# 4 Decision Tree

## 4.1 a

Conditional Entropy of accident rate (Y) when weather(X) is chosen as first predictor variable to split:
$\rightarrow H(Y/X) = -(\frac{1}{2}(\frac{1}{2}\log\frac{1}{2}) + (\frac{1}{2}\frac{1}{2}\log\frac{1}{2}))$
$\rightarrow 1$
With Traffic chosen as predictor variable to split first:
$\rightarrow H(Y/X) = -(\frac{1}{2}(\frac{2}{2}\log\frac{2}{2}) + \frac{1}{2}(\frac{2}{2}\log\frac{2}{2}))$
$\rightarrow 0$
Information Gain with Traffic is more. Hence, we choose Traffic as the first predictor variable.

## 4.2 b

The trees $T_1, T_2$ will have identical structure i,e, same pattern of predictor variables and leaf nodes.
Since probabilities are used while calculating entropies, normalization should not change the tree structure.

## 4.3 c

$p_k <= \frac{2^{p_k}}{2}$ for $p_k$ in [0,1]
$\rightarrow \log p_k <= p_k - 1$
$\rightarrow 1 - p_k + \log p_k <= 0$
$\rightarrow p_k(1 - p_k + \log p_k) <= 0$
$\rightarrow \sum_{k=1}^{K} 1 - p_k + \log p_k <= 0$
$\rightarrow \sum_{k=1}^{K} p_k(1 - p_k) <= \sum_{k=1}^{K} p_k(\log p_k)$
$\rightarrow GiniIndex <= Cross - entropy$

# 5 Programming

## 5.1 Data Inspection

There are 11 attributes.
Not all of them are meaningful for classification.
The first attribute ID number is not meaningful attribute for classification. The last attribute (class) is not a feature attribute

| k | L1 Training Accuracy | L2 Training Accuracy | L1 Testing Accuracy | L2 Testing Accuracy |
|---|---|---|---|---|
| 1 | 75.00 | 71.43 | 66.67 | 66.67 |
| 3 | 72.96 | 71.43 | 72.22 | 66.67 |
| 5 | 70.92 | 70.92 | 66.67 | 66.67 |
| 7 | 69.90 | 69.39 | 72.22 | 66.67 |

Table 1: kNN model

| Training Accuracy | Testing Accuracy |
|---|---|
| 45.92 | 33.33 |

Table 2: Naive Bayes model

There are 7 classes.
The classes do not follow uniform distribution. Class 2 has the highest occurrence followed by class 1 in the training set. Class 4 has no occurrence.

### 5.2 Performance Comparison

kNN results are shown in Table 1
Naive bayes results are shown in Table 2
The classification results from the kNN Classifier are better than the results from Naive Bayes Classifier.
This may be because the conditional independence assumption used for the Naive bayes model is not accurate with the given features. There may be dependency among the features.