1 Density Estimation

(a)
- The PDF of Beta distribution is

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

When $\beta=1$, it becomes

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

$$L=L(\alpha)=\prod_{i=1}^{n}f(x_i;\alpha)=\alpha^n\prod_{i=1}^{n}x_i^{\alpha-1}$$

$$\ln L=n\ln\alpha+(\alpha-1)\sum_{i=1}^{n}\ln x_i$$

Let

$$\frac{\partial\ln L}{\partial\alpha}=\frac{n}{\alpha}+\sum_{i=1}^{n}\ln x_i=0$$

We have

$$\alpha=\frac{n}{-\sum_{i=1}^{n}\ln x_i}$$

-

$$f(x;\theta)=N(\theta,\theta)=\frac{1}{\sqrt{2\pi\theta}}e^{-\frac{(x-\theta)^2}{2\theta}}$$

$$L=L(\theta)=\prod_{i=1}^{n}f(x_i,\theta)=(2\pi\theta)^{-\frac{n}{2}}\prod_{i=1}^{n}e^{-\frac{(x_i-\theta)^2}{2\theta}}$$

$$\ln L=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\theta)-\frac{1}{2\theta}\sum_{i=1}^{n}(x_i-\theta)^2=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\theta)-\frac{1}{2\theta}\sum_{i=1}^{n}x_i^2+\sum_{i=1}^{n}x_i-\frac{n\theta}{2}$$

Let

$$\frac{\partial\ln L}{\partial\theta}=-\frac{n}{2\theta}+\frac{\sum_{i=1}^{n}x_i^2}{2\theta^2}-\frac{n}{2}=0$$

Then we have

$$\frac{\sum_{i=1}^{n}x_i^2}{\theta^2}-\frac{n}{\theta}-n=0$$

We know $\theta>0$, so

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

(b)

- $$\begin{aligned}\mathbf{E}_{X_1,\cdots,X_n}[\hat{f}(x)] &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_{X_i}\left[\frac{1}{h}K\left(\frac{x-X_i}{h}\right)\right]\\ &= \mathbf{E}\left[\frac{1}{h}K\left(\frac{x-X}{h}\right)\right]\\ &= \int\frac{1}{h}K(\frac{x-t}{h})f(t)dt = \frac{1}{h}\int K(\frac{x-t}{h})f(t)dt\end{aligned}$$

- By letting $z = \frac{x-t}{h}$, we have

  $$\frac{1}{h}\int K(\frac{x-t}{h})f(t)dt = \int K(z)f(x-zh)dz$$

  By Taylor's theorem, we have

  $$f(x-zh) = f(x) - zhf'(x) + \frac{z^2h^2}{2}f''(x) + \cdots$$

- $$\begin{aligned}\mathbf{E}[\hat{f}(x)] = \int K(z)f(x-zh)dz &= \int K(z)dz\left\{f(x) - zhf'(x) + \frac{z^2h^2}{2}f''(x) + \cdots\right\}\\ &= f(x)\int K(z)dz - f'(x)h\int zK(z)dz + f''(x)\frac{h^2}{2}\int z^2K(z)dz + \cdots\\ &= f(x) + \frac{h^2\sigma_K^2}{2}f''(x) + o(h^2)\end{aligned}$$

  So we know the bias term is

  $$\mathbf{E}[\hat{f}(x)] - f(x) = \frac{h^2\sigma_K^2}{2}f''(x) + o(h^2)$$

2 Histogram Density Estimates

Part (a)

- The estimate of density is equivalent to the portion of the samples that have fallen within the given bin devided by the length of bin(h). Mathematically:

  $$\hat{f_n})(x) = \frac{1}{n*h}\sum_{i=1}^{N}\mathbf{1}_{(x_0,x_0+h]}(x_i)$$

  where $\mathbf{1}_{(x_0,x_0+h]}(x_i)$ is 1 if $x_i \in (x_0, x_0+h]$ and 0 otherwise.

- Mean of B is np hence:

  $$E(\hat{f_n})(x)) = \frac{1}{nh}E(B) = \frac{F(x_0+h)-F(x_0)}{h}$$

- Var of B is $np(1-p)$ hence:

$$Var(\hat{f_n})(x)) = \frac{1}{n^2h^2}E(B) = E(\hat{f_n}(x))\frac{1-F(x_0+h)+F(x_0)}{nh}$$

- if we let $h \to 0$ and $n \to \infty$ then:

$$E(\hat{f_n}(x)) \to f(x_n)$$

since the pdf is the derivative of the CDF. But since x is between x_0 and x_0+h, $f(x_0) \to f(x)$. So if we use smaller and smaller bins as we get more data, the histogram density estimate is unbiased. Wed also like its variance to shrink as the same grows. Since $1-F(x_0+h)+F(x_0) \to 1$ as $h \to 0$, to get the variance to go away we need $nh \to \infty$.

To put this together, then, our first conclusion is that histogram density estimates will be consistent when $h \to 0$ but $nh \to \infty$ as $n \to \infty$ The bin-width h needs to shrink, but slower than n^{-1}.

3 Naive Bayes

(a) (**10 points**)Suppose $X = \{X_i\}_{i=1}^D \in \mathbb{R}^D$ represents the features and $Y \in \{0,1\}$ represents the class labels. Let the following assumptions hold:

(a) The label variable Y follows a Bernoulli distribution, with parameter $\boldsymbol{\pi} = P(Y=1)$.

(b) For each feature X_j, we have $P(X_j|Y=y_k)$ which follows a Gaussian distribution $N(\mu_{jk}, \sigma_j)$.

Using the Naive Bayes assumption, *"for all $j' \neq j$, X_j and $X_{j'}$ are conditionally independent given Y"*, compute $P(Y=1|X)$ and show that it can be written in the following form:

$$P(Y=1|X) = \frac{1}{1+\exp(-w_0+\boldsymbol{w}^T\boldsymbol{X})}$$

Specifically, find the explicit form of w_0 and $\boldsymbol{w}$ in terms of $\boldsymbol{\pi}, \mu_{jk}$, and σ_j, for $j = 1, \ldots, D$, and $k \in \{0,1\}$. **Solution:** For ease of indexing, and without loss of generality, let $y_0 = 0$ and $y_1 = 1$:

$$P(Y=1|X) = P(X|Y=1)\frac{P(Y=1)}{P(X)} \quad = \pi\frac{P(X|Y=1)}{P(X)}.$$

with

$$P(Y=1) = \pi$$

Now

$$P(X) = \sum_y P(X|Y=y) = \pi P(X|Y=1) + (1-\pi)P(X|Y=0)$$

Thus

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

Explicitly, using

$$P(X|Y=y_k) = \prod_{j=1}^{D} (2\pi\sigma_j^2)^{-\frac{1}{2}} \exp(-(2\sigma_j^2)^{-1}(x_j - \mu_{jk})^2)$$

We get

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

Consider only the second term of the denominator:

$$\begin{aligned}
&(\frac{1}{\boldsymbol{\pi}} - 1)\prod_{j=1}^{D} \exp((2\sigma_j^2)^{-1}((x_j - \mu_{j1})^2 - (x_j - \mu_{j0})^2) \\
&= \exp(\log(\frac{1}{\boldsymbol{\pi}} - 1)) \exp(\sum_j (2\sigma_j^2)^{-1}(x_j^2 - 2x_j\mu_{j1} + \mu_{j1}^2 - x_j^2 + 2x_j\mu_{j0} - \mu_{j2}^2)) \\
&= \exp(\log(\frac{1}{\boldsymbol{\pi}} - 1) + \sum_j (2\sigma_j^2)^{-1}(\mu_{j1}^2 - \mu_{j2}^2) + \sum_j (\sigma_j^2)^{-1}(\mu_{j0} - \mu_{j1})x_j)
\end{aligned}$$

Comparing this with the second term in the denominator of

$$P(Y=1|X) = \frac{1}{1 + \exp(-w_0 + \boldsymbol{w}^T\boldsymbol{X})}$$

gives us:

(a) $w_0 = -\log(\frac{1}{\boldsymbol{\pi}} - 1) - \sum_j \left[(2\sigma^2)^{-1}(\mu_{j1}^2 - \mu_{j0}^2)\right].$

(b) $w_j = \frac{\mu_{j0} - \mu_{j1}}{\sigma_j^2}, for j >= 1.$

(b) **(10 points) IMPORTANT: if someone solves the problem with σ_{jk} or σ_j, either one is fine. Here is two solutions.**
Case 1: σ_{jk}

The data $\mathcal{D} = \{(X_i, Y_i\}_{i=1}^{N}$, the parameters $\Theta = \{p_k, \mu_{jk}, \sigma_{jk}\}_{j=1,\cdots,D; k=1,\cdots,K}$.

By Bayes rules, we know

$$\begin{aligned}
P(X_i, Y_i; \Theta) &= P(Y_i)P(X_i|Y_i) = P(Y_i)\prod_{j=1}^{D} P(x_{ij}|Y_i) \\
&= p_{Y_i}\prod_{j=1}^{D} \mathcal{N}(\mu_{j,Y_i}, \sigma_{j,Y_i}) = p_{Y_i}\prod_{j=1}^{D} \frac{1}{\sqrt{2\pi\sigma_{j,Y_i}}} e^{-\frac{(x_{ij} - \mu_{j,Y_i})^2}{2\sigma_{j,Y_i}}}
\end{aligned}$$

The log likelihood is

$$\begin{aligned}
L = L(\mathcal{D};\Theta) &= \ln \prod_{i=1}^{N} P(X_i, Y_i; \Theta) = \sum_{i=1}^{N} \ln P(X_i, Y_i; \Theta) \\
&= \sum_{i=1}^{N} \ln \left(p_{Y_i} \prod_{j=1}^{D} \frac{1}{\sqrt{2\pi\sigma_{j,Y_i}}} e^{-\frac{(x_{ij}-\mu_{j,Y_i})^2}{2\sigma_{j,Y_i}}} \right) \\
&= \sum_{i=1}^{N} \ln p_{Y_i} + \sum_{i=1}^{N}\sum_{j=1}^{D} \ln \frac{1}{\sqrt{2\pi\sigma_{j,Y_i}}} - \sum_{i=1}^{N}\sum_{j=1}^{D} \frac{(x_{ij}-\mu_{j,Y_i})^2}{2\sigma_{j,Y_i}} \\
&= \sum_{i=1}^{N} \ln p_{Y_i} - \frac{ND\ln 2\pi}{2} - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{D} \sigma_{j,Y_i} - \sum_{i=1}^{N}\sum_{j=1}^{D} \frac{(x_{ij}-\mu_{j,Y_i})^2}{2\sigma_{j,Y_i}}
\end{aligned}$$

To solve the MLE problem with constrain $\sum_{k=1}^{K} p_k = 1$, we use Lagrange multiplier:

$$\mathcal{L} = L + \lambda(\sum_{k=1}^{K} p_k - 1)$$

We also denote $N_k = \sum_{i=1}^{N} \mathbb{I}(Y_i = k)$ as the number of samples belonging to class k. We let

$$\frac{\partial \mathcal{L}}{\partial p_k} = \sum_{i=1}^{N} \frac{\mathbb{I}(Y_i = k)}{p_{Y_i}} + \lambda = \frac{N_k}{p_{Y_i}} + \lambda = 0$$

Then we have

$$p_k = \frac{N_k}{-\lambda} = \frac{N_k}{N}$$

Also, we let

$$\frac{\partial \mathcal{L}}{\partial \mu_{jk}} = \sum_{i=1}^{N} \mathbb{I}(Y_i = k)\frac{(x_{ij}-\mu_{jk})}{\sigma_{jk}} = \frac{1}{\sigma_{jk}}\sum_{i=1}^{N} \mathbb{I}(Y_i = k)(x_{ij}-\mu_{jk}) = 0$$

Then we have

$$\mu_{jk} = \frac{\sum_{i=1}^{N} \mathbb{I}(Y_i = k)x_{ij}}{N_k}$$

Also, we let

$$\frac{\partial \mathcal{L}}{\partial \sigma_{jk}} = -\frac{1}{2}\sum_{i=1}^{N} \mathbb{I}(Y_i = k)\sigma_{jk} + \sum_{i=1}^{N} \mathbb{I}(Y_i = k)\frac{(x_{ij}-\mu_{jk})^2}{2\sigma_{jk}^2} = -\frac{N_k}{2\sigma_{jk}} + \frac{\sum_{i=1}^{N} \mathbb{I}(Y_i = k)(x_{ij}-\mu_{jk})^2}{2\sigma_{jk}^2} = 0$$

Since $\sigma_{jk} > 0$, we have

$$\sigma_{jk} = \frac{\sum_{i=1}^{N} \mathbb{I}(Y_i = k)(x_{ij}-\mu_{jk})^2}{N_k} = \frac{\sum_{i=1}^{N} \mathbb{I}(Y_i = k)(x_{ij} - \frac{\sum_{i=1}^{N}\mathbb{I}(Y_i=k)x_{ij}}{N_k})^2}{N_k}$$

Case 2. It has same answers for μ and π σ_j
The data $\{(x_i, y_i)\}_{i=1}^N$, the parameters $\Theta = \{\pi_k, \mu_{jk}, \sigma_j\}_{j=1,\cdots,D;k\in 0,1}$

The log likelihood is

$$
\begin{aligned}
L = L(\mathcal{D};\Theta) &= \ln \prod_{i=1}^N P(X_i, Y_i;\Theta) = \sum_{i=1}^N \ln P(X_i, Y_i;\Theta) \\
&= \sum_{i=1}^N \ln \left(\pi_{y_i} \prod_{j=1}^D \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_{ij}-\mu_{j,y_j})^2}{2\sigma_j^2}} \right) \\
&= \sum_{i=1}^N \ln \pi_{y_i} + \sum_{i=1}^N \sum_{j=1}^D \ln \frac{1}{\sqrt{2\pi\sigma_j^2}} - \sum_{i=1}^N \sum_{j=1}^D \frac{(x_{ij}-\mu_{j,y_i})^2}{2\sigma_j^2} \\
&= \sum_{i=1}^N \ln \pi_{y_i} - \frac{ND\ln 2\pi}{2} - \sum_{i=1}^N \sum_{j=1}^D \ln \sigma_j \\
&- \sum_{i=1}^N \sum_{j=1}^D \frac{(x_{ij}-\mu_{j,y_i})^2}{2\sigma_j^2}
\end{aligned}
$$

We let

$$
L(\sigma_j) = -\sum_{i=1}^N \sum_{j=1}^D \ln \sigma_j - \sum_{i=1}^N \sum_{j=1}^D \frac{(x_{ij}-\mu_{j,y_i})^2}{2\sigma_j^2}
$$

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \sigma_j^2} &= -\frac{N}{2\sigma_j^2} + \sum_{i=1}^N \mathbb{I}(Y_i = k)\frac{(x_{ij}-\mu_{jk})^2}{2\sigma_j^4} \\
&= -\frac{N}{2\sigma_j^2} + \frac{\sum_{i=1}^N \mathbb{I}(Y_i = k)(x_{ij}-\mu_{jk})^2}{2\sigma_j^4} = 0
\end{aligned}
$$

Since $\sigma_{jk} > 0$, we have

$$
\sigma_j^2 = \frac{\sum_{i=1}^N \mathbb{I}(Y_i = k)(x_{ij}-\mu_{jk})^2}{N}
$$

4 Nearest Neighbor

(a) Let us follow the following label convention:

- Mathematics: 1
- Electrical Engineering: 2
- Computer Science: 3
- Economics : 4

Thus the unnormalized data is:

x-coordinate	y-coordinate	label
0	49	1
-7	32	1
-9	47	1
29	12	2
49	31	2
37	38	2
8	9	3
13	-1	3
-6	-3	3
-21	12	3
27	-32	4
19	-14	4
27	-20	4

Table 1: Unnormalized labeled data.

The mean and standard deviation are:

	x-coordinate	y-coordinate
mean	12.77	12.31
std	20.7170	25.9306

Table 2: Mean and standard deviation of the labeled data.

The normalized queried student coordinate is:

normalized queried student x-coordinate	normalized queried student y-coordinate
0.3490	-0.2047

Table 3: Normalized queried data.

The L_1 and L_2 distances between the queried student coordinate and each labeled data are:

L_1 Distance	L_2 Distance	label
2.5851	1.8856	1
2.2674	1.6211	1
2.9424	2.0830	1
0.6272	0.4753	2
2.3253	1.6781	2
2.0161	1.4500	2
0.6564	0.5843	3
0.6464	0.4575	3
1.6407	1.3129	3
2.1719	1.9884	4
1.8419	1.5415	4
0.8581	0.8113	4
1.3791	1.0947	4

Table 4: L_1 and L_2 distances between normalized queried data and each of the normalized labeled data.

If sorted (from minimum to maximum) by L_2 distance:

L_2 Distance	label
0.4575	3
0.4753	2
0.5843	3
0.8113	4
1.0947	4
1.3129	3
1.4500	2
1.5415	4
1.6211	1
1.6781	2
1.8856	1
1.9884	3
2.0830	1

Table 5: Sorting based on L_2 distance between normalized queried data and each of the normalized labeled data.

If sorted (from minimum to maximum) by L_1 distance:

L_1 Distance	label
0.6272	2
0.6464	3
0.6564	3
0.8581	4
1.3791	4
1.6407	3
1.8419	4
2.0161	2
2.1719	3
2.2674	1
2.3254	2
2.5851	1
2.9424	1

Table 6: Sorting based on L_1 distance between normalized queried data and each of the normalized labeled data.

Thus:

- If using L_2 distance metric and $K = 1$, the predicted student major will be label 3 (**Computer Science**).
- If using L_2 distance metric and $K = 5$, the predicted student major will be label 2 (**Computer Science**). Actually this is a tie between Computer Science and Economics, but tie-breaking is chosen by the labeled data with shortest distance.
- If using L_1 distance metric and $K = 1$, the predicted student major will be label 3 (**Electrical Engineering**).
- If using L_1 distance metric and $K = 5$, the predicted student major will be label 3 (**Computer Science**). Actually this is a tie between Computer Science and Economics, but tie-breaking is chosen by the labeled data with shortest distance.

(b) Probabilistic K-Nearest Neighbor:

- The unconditional density $p(\mathbf{x})$ can be computed as follows:

$$\begin{aligned} p(\mathbf{x}) &= \sum_c p(\mathbf{x} \mid Y = c)\, p(Y = c) \\ &= \sum_c \frac{K_c}{N_c V} \frac{N_c}{N} \\ &= \sum_c \frac{K_c}{NV} \\ p(\mathbf{x}) &= \frac{K}{NV} \end{aligned}$$

- The posterior probability of class membership $p\left(Y=c \mid \mathbf{x}\right)$ can be computed as follows:

$$\begin{aligned} p\left(Y=c \mid \mathbf{x}\right) &= \frac{p\left(\mathbf{x} \mid Y=c\right) p\left(Y=c\right)}{p\left(\mathbf{x}\right)} \\ &= \frac{\frac{K_c}{N_c V}\frac{N_c}{N}}{\frac{K}{NV}} \\ &= \frac{\frac{K_c}{NV}}{\frac{K}{NV}} \\ p\left(Y=c \mid \mathbf{x}\right) &= \frac{K_c}{K} \end{aligned}$$

5 MLE and MAP

(a) (**3 points**)

(a) The joint probability distribution, $P(X=x, P=p)$

$$\begin{aligned} P(X=x, P=p) &= P(X=x \mid P=p)P(P=p) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \mathbb{I}(0<p<1) \\ &= \binom{n}{x} p^x (1-p)^{n-x} \end{aligned}$$

(b) The marginal probability distribution, $P(X=x)$

$$\begin{aligned} P(X) &= \int_0^1 P(X,p)dp \\ &= \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} dp \\ &= \binom{n}{x} B(x+1, n-x+1) \end{aligned}$$

(c) The posterior distribution, $P(P=p \mid X=x)$

$$\begin{aligned} P(P=p \mid X=x) &= \frac{P(P=p, X=x)}{P(X=x)} \\ &= \frac{\binom{n}{x} p^x (1-p)^{n-x}}{\binom{n}{x} B(x+1, n-x+1)} \\ &= \frac{p^x (1-p)^{n-x}}{B(x+1, n-x+1)} \end{aligned}$$

(b) (**3 points**)

(a) The marginal probability distribution, $P(X = x)$

$$\begin{aligned}
P(X = x) &= \int_0^1 P(X, p)dp \\
&= \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} dp \\
&= \frac{\binom{n}{x}}{B(\alpha,\beta)} \int_0^1 p^{x+\alpha-1}(1-p)^{n-x+\beta-1} dp \\
&= \frac{\binom{n}{x} B(x+\alpha, n-x+\beta)}{B(\alpha,\beta)}
\end{aligned}$$

(b) The posterior distribution, $P(P = p \mid X = x)$

$$\begin{aligned}
P(P = p \mid X = x) &= \frac{P(P = p, X = x)}{P(X = x)} \\
&= \frac{\binom{n}{x} p^x (1-p)^{n-x} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}}{\frac{\binom{n}{x} B(x+\alpha,n-x+\beta)}{B(\alpha,\beta)}} \\
&= \frac{p^{x+\alpha-1}(1-p)^{n-x+\beta-1}}{B(x+\alpha, n-x+\beta)}
\end{aligned}$$

(c) (**9 points**)

(a) MLE and MAP of (a)

$$\begin{aligned}
\frac{\partial P(X = x \mid P = p)}{\partial p} &= \frac{\partial \binom{n}{x} p^x (1-p)^{n-x}}{\partial p} \\
&= \binom{n}{x} x p^{x-1}(1-p)^{n-x} - p^x(n-x)(1-p)^{n-x-1} \\
&= \binom{n}{x} p^{x-1}(1-p)^{n-x-1}\left(x(1-p) - (n-x)p\right) \\
\therefore p &= \frac{x}{n}
\end{aligned}$$

MAP has a same result because the prior is independent on p. If prior is independent on the paramter p, the estimates of MLE and MAP are same.

(b) MLE and MAP of (b)
Estimate of MLE is same as above.

$$\begin{aligned}
&\frac{\partial P(P = p \mid X = x)}{\partial p} = \frac{1}{B(x+\alpha, n-x+\beta)} \frac{\partial p^{x+\alpha-1}(1-p)^{n-x+\beta-1}}{\partial p} \\
&= \frac{1}{B(x+\alpha, n-x+\beta)} \left((x+\alpha-1)p^{x+\alpha-2}(1-p)^{n-x+\beta-1} - p^{x+\alpha-1}(n-x+\beta-1)(1-p)^{n-x+\beta-2}\right) \\
&= \frac{1}{B(x+\alpha, n-x+\beta)} p^{x+\alpha-2}(1-p)^{n-x+\beta-2}\left((x+\alpha-1)(1-p) - p(n-x+\beta-1)\right) \\
&\therefore p = \frac{x+\alpha-1}{n+\alpha+\beta-2}
\end{aligned}$$

Estimates of MLE and MAP are different because of its prior distribution.
When $x = 2, n = 10$, we will say that $p = 0.2$ under the MLE estimation. However, it is going to be $p = \frac{2+50-1}{10+50+50-2} = 0.4722$. If we have a probable prior distribution as like $\alpha = 50, \beta = 50$ (i.e., a coin is fair.), the MAP estimation is not sensitive on the small number of exceptional occurrences(2 out of 10). Thus, MAP is more robust than MLE.

6 Decision Tree

Part (a) We should split on "Traffic" because it gives a perfect prediction of "Accident rate". The other cannot do perfect prediction. 5 Just mentioning the fact that Traffic gives perfect prediction.

Part (b) We can think about decision trees as partitioning the space of observations along each axis. If every feature is continuous and ordered we can transform T_1 into T_2 by taking each decision boundary, subtracting off the appropriate mean, and then dividing by the appropriate variance. Both trees have the same structure and same accuracy. In other words, linear transformation does not change informativeness of the features. 5 The argument that informativeness doesn't change with linear transformation.

Part (c)
Consider the difference between the Gini Index and Cross Entropy:

$$G - CE = \sum_{k=1}^{K}[p_k(1-p_k)] + \sum_{k=1}^{K}[p_k \log p_k]$$

$$G - CE = \sum_{k=1}^{K} p_k(1 - p_k + \log p_k)$$

Now examine the function $f(x) = 1 - x + \log(x)$, where the base of the log is less than or equal to e (the cross entropy is defined with base 2). Note that f is continuous on the positive real line.

Now consider the derivative $\frac{d}{dx}f = -1 + \frac{1}{x\log(a)}$ where a is the base of the log. This function is also continuous on the positive real line. For all $a \leq e$, $\log(a) \leq 1 \Rightarrow \frac{1}{x\log(a)} \leq 1$ for all $x \in (0,1)$, and for $x = 1$, $\frac{1}{x\log(a)} = 1$. This implies that $\frac{d}{dx}f(x) > 0$ for $x \in (0,1), a < e$ so f has no critical points in $(0,1)$.

Note that $f(x) \to -\infty$ as $x \to 0+$ and consider $x = 1$. $f(x) = 0$, and has no previous critical points, so it cannot have any positive points (if f were to have a positive point, since it is continous it must decrease to $f(0)$, but it then must have a negative derivative, meaning its derivative must have a zero, meaning it must have a critical point. Contradiction.).

Thus, $1 - p_k + \log p_k < 0$, meaning that $G - CE < 0$, meaning that the Gini Index is always less than the Cross Entropy. 5 Any correct proof is acceptable. Some partial credit should be given as well if some of the ideas are correct.