CIS 520, Machine Learning, Fall 2018: Assignment 1

Shubhankar Patankar

September 30, 2018

Collaborators:

1 Conditional independence in probability models

1. $p(x_i) = \sum_{j=1}^{k} f_j(x_i)\pi_j$

2. The formula for $p(x_1, \dots, x_n)$ can be derived as follows:
   Assuming $x_1, \dots, x_n$ are independent,

$$p(x_1, \dots, x_n) = \prod_{m=1}^{n} p(x_m)$$

$$\therefore p(x_1, \dots, x_n) = \prod_{m=1}^{n} \left(\sum_{j=1}^{k} f_j(x_m)\pi_j \right)$$

3. The formula for $p(z_u = v \mid x_1, \dots, x_n)$ can be derived as follows:
   It is known that $p(x_u \mid z_u = v) = f_v(x_u)$. Therefore, the probability of the u-th data point $p(x_u)$ can be uniquely determined given the knowledge that it is generated by function v. Similarly, the probability that the u-th data point is generated by the v-th function is dependent solely on the knowledge of the data point x_u itself, not the entire data set.

$$\therefore p(z_u = v \mid x_1, \dots, x_n) = p(z_u = v \mid x_u)$$

   Using Bayes Rule,

$$p(z_u = v \mid x_u) = \frac{p(x_u \mid z_u = v)p(z_u = v)}{p(x_u)}$$

$$\therefore p(z_u = v \mid x_u) = \frac{f_v(x_u)\pi_v}{\sum_{j=1}^{k} f_j(x_u)\pi_j}$$

   Alternatively,

$$p(z_u = v \mid x_1, \dots, x_n) = \frac{p(x_1, \dots, x_n \mid z_u = v)p(z_u = v)}{p(x_1, \dots, x_n)}$$

$$p(z_u = v \mid x_1, \dots, x_n) = \frac{p(x_u \mid z_u = v)p(z_u = v)\left[\prod_{i=1, i\neq u}^{n} \left(\sum_{j=1}^{k} f_j(x_i)\pi_j\right)\right]}{\sum_{j=1}^{k} f_j(x_u)\pi_j \left[\prod_{i=1, i\neq u}^{n} \left(\sum_{j=1}^{k} f_j(x_i)\pi_j\right)\right]}$$

$$\therefore p(z_u = v \mid x_u) = \frac{f_v(x_u)\pi_v}{\sum_{j=1}^{k} f_j(x_u)\pi_j}$$

2 Non-Normal Norms

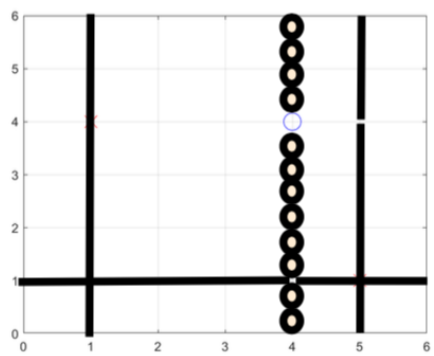
1. For the given vectors,the point closest to x_1 under each of the following norms is

   $\|x_1 - x_2\| = [0.1, -0.6, -0.3, -0.4]$
   $\|x_1 - x_3\| = [0.2, -0.9, 0.1, 0]$
   $\|x_1 - x_4\| = [0, 2.6, 0, 0.9]$

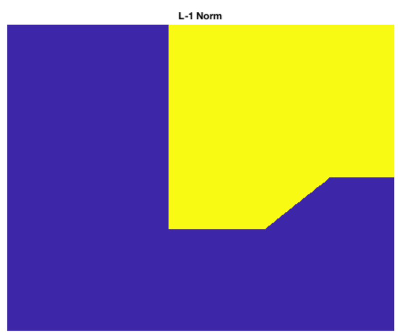
   a) L_0: x_4 with distance = 2
   b) L_1: x_3 with distance = 1.2
   c) L_2: x_2 with distance = 0.787
   d) L_{inf}: x_2 with distance = 0.6

2. Draw the 1-Nearest Neighbor decision boundaries with the given norms and lightly shade the o region: In all figures below, blue refers to 'x' and yellow to 'o'. For the L-0 norm, most of the space is unclassifiable with the exception of the lines marked. Points (4,1) and (5,4) are not classifiable. The code used to generate decision boundaries for norms L_1, L_2 and L_{inf} is included in the Appendix.

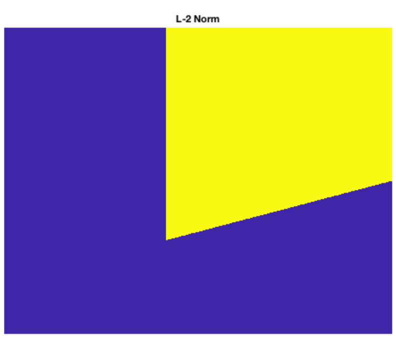
   a) L_0



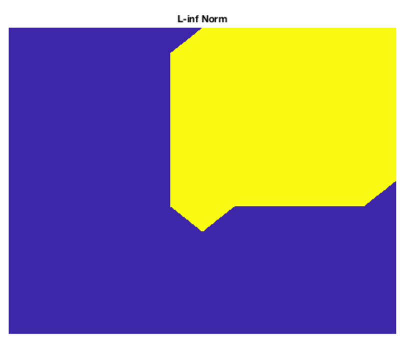
   b) L_1



   c) L_2



   d) L_{inf}



3 Decision trees

3.1 Part 1

The sample entropy $H(Y)$ can be written as:

$$H(Y) = -P(Y=+)\log_2(P(Y=+)) - P(Y=-)\log_2(P(Y=-))$$

$$\therefore H(Y) = (16/30)\log_2(16/30) - (14/30)\log_2(14/30) = 0.9968$$

The information gains are given by $IG(X_i) = H(Y) - H(Y \mid X_i)$.

$$\begin{aligned}
\therefore H(Y \mid X_1) = P(X_1 = T)\Big[& -P(Y=+\mid X_1=T)\log_2(-P(Y=+\mid X_1=T)) \\
& - P(Y=-\mid X_1=T)\log_2(-P(Y=-\mid X_1=T))\Big] \\
& + P(X_1=F)\Big[-P(Y=+\mid X_1=F)\log_2(-P(Y=+\mid X_1=F)) \\
& - P(Y=-\mid X_1=F)\log_2(-P(Y=-\mid X_1=F))\Big] \\
& = (13/30)\Big[(-6/13)\log_2(6/13) - (7/13)\log_2(7/13)\Big] \\
& + (17/30)\Big[(-10/17)\log_2(10/17) - (7/17)\log_2(7/17)\Big] \\
& = 0.9852
\end{aligned}$$

$$\therefore IG(X_1) = 0.9968 - 0.9852 = 0.0115$$

$$\begin{aligned}
H(Y \mid X_2) = P(X_2 = T)\Big[& -P(Y=+\mid X_2=T)\log_2(-P(Y=+\mid X_2=T)) \\
& - P(Y=-\mid X_2=T)\log_2(-P(Y=-\mid X_2=T))\Big] \\
& + P(X_2=F)\Big[-P(Y=+\mid X_2=F)\log_2(-P(Y=+\mid X_2=F)) \\
& - P(Y=-\mid X_2=F)\log_2(-P(Y=-\mid X_2=F))\Big] \\
& = (11/30)\Big[(-4/11)\log_2(4/11) - (7/11)\log_2(7/11)\Big] \\
& + (19/30)\Big[(-12/19)\log_2(12/19) - (7/19)\log_2(7/19)\Big] \\
& = 0.9480
\end{aligned}$$

$$\therefore IG(X_2) = 0.9968 - 0.9480 = 0.0488$$

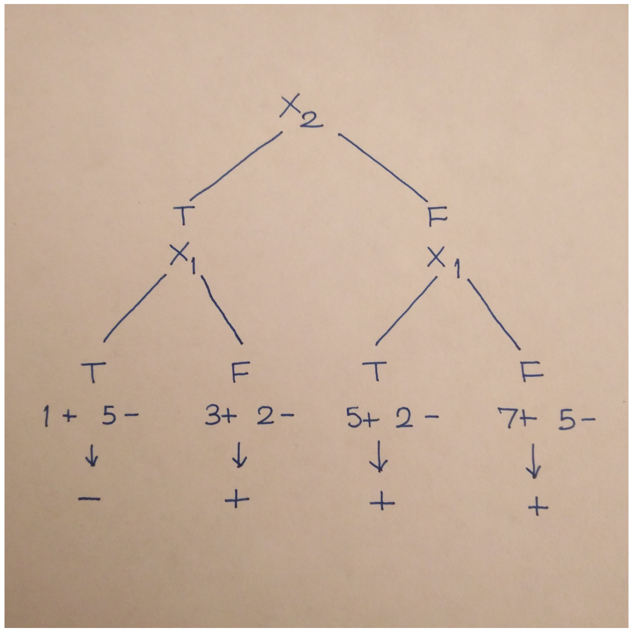


Figure 1: Decision Tree

3.2 Part 2

1. If variables X and Y are independent, is $IG(x, y) = 0$? If yes, prove it. If no, give a counter example.

$$X \perp Y \implies p(x, y) = p(x)p(y)$$

$$\begin{aligned}
\therefore IG(x, y) &= -\sum_x \sum_y p(x, y) \log \left(\frac{p(x)p(y)}{p(x, y)} \right) \\
&= -\sum_x \sum_y p(x, y) \log \left(\frac{p(x)p(y)}{p(x)p(y)} \right) \\
&= -\sum_x \sum_y p(x, y) \log (1) = 0
\end{aligned}$$

2. Proof that $IG(x,y) = H[x] - H[x \mid y] = H[y] - H[y \mid x]$, starting from the definition in terms of KL-divergence:

$$\begin{aligned}
IG(x,y) &= KL\left(p(x,y) || p(x)p(y)\right) \\
&= -\sum_x \sum_y p(x,y) \log\left(\frac{p(x)p(y)}{p(x,y)}\right) \\
&= -\sum_x \sum_y p(x,y) \big[\log p(x) + \log p(y) - \log p(x,y)\big] \\
&= -\sum_x \sum_y \big[p(x,y)\log p(x) + p(x,y)\log p(y) - p(x,y)\log p(x,y)\big] \\
&= -\sum_x \sum_y p(x,y)\log p(x) - \sum_x \sum_y p(x,y)\log p(y) + \sum_x \sum_y p(x,y)\log p(x,y) \\
&= -\sum_x p(x)\log p(x) - \sum_y p(y)\log p(y) + \sum_x \sum_y p(x,y)\log p(x,y) \\
&= H[x] + H[y] - H[x,y] \\
&= H[x] + H[y] - (H[y \mid x] + H[x]) \\
&= H[y] - H[y \mid x] \\
&= H[x] - H[x \mid y]
\end{aligned}$$

4 High dimensional hi-jinx

1. Intra-class distance.

$$\begin{aligned}
\mathbf{E}[(X - X')^2] &= \mathbf{E}[X^2 - 2XX' + X'^2] \\
&= \mathbf{E}[X^2] - 2\mathbf{E}[XX'] + \mathbf{E}[X'^2] \\
&= \mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X'] + \mathbf{E}[X'^2] \\
&= ({\mu_1}^2 + \sigma^2) - 2{\mu_1}^2 + ({\mu_1}^2 + \sigma^2) \\
&= 2\sigma^2
\end{aligned}$$

2. Inter-class distance.

$$\begin{aligned}
\mathbf{E}[(X - X')^2] &= \mathbf{E}[X^2 - 2XX' + X'^2] \\
&= \mathbf{E}[X^2] - 2\mathbf{E}[XX'] + \mathbf{E}[X'^2] \\
&= \mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X'] + \mathbf{E}[X'^2] \\
&= ({\mu_1}^2 + \sigma^2) - 2\mu_1\mu_2 + ({\mu_2}^2 + \sigma^2) \\
&= {\mu_1}^2 - 2\mu_1\mu_2 + {\mu_2}^2 + 2\sigma^2 \\
&= (\mu_1 - \mu_2)^2 + 2\sigma^2
\end{aligned}$$

3. Intra-class distance, m-dimensions.

$$
\begin{aligned}
\mathbf{E}\Big[\sum_{j=1}^{m}(X_j - X_j')^2\Big] &= \mathbf{E}\Big[\sum_{j=1}^{m}\big({X_j}^2 - 2X_jX_j' + {X_j'}^2\big)\Big] \\
&= \mathbf{E}\Big[\sum_{j=1}^{m}{X_j}^2\Big] - 2\mathbf{E}\Big[\sum_{j=1}^{m}X_jX_j'\Big] + \mathbf{E}\Big[\sum_{j=1}^{m}{X_j'}^2\Big] \\
&= \sum_{j=1}^{m}\mathbf{E}\big[{X_j}^2\big] - 2\sum_{j=1}^{m}\mathbf{E}\big[X_jX_j'\big] + \sum_{j=1}^{m}\mathbf{E}\big[{X_j'}^2\big] \\
&= \sum_{j=1}^{m}\mathbf{E}\big[{X_j}^2\big] - 2\sum_{j=1}^{m}\mathbf{E}\big[X_j\big]\mathbf{E}\big[X_j'\big] + \sum_{j=1}^{m}\mathbf{E}\big[{X_j'}^2\big] \\
&= \sum_{j=1}^{m}({\mu_{1j}}^2 + \sigma^2) - 2\sum_{j=1}^{m}{\mu_{1j}}^2 + \sum_{j=1}^{m}({\mu_{1j}}^2 + \sigma^2) \\
&= m\sigma^2 + m\sigma^2 = 2m\sigma^2
\end{aligned}
$$

4. Inter-class distance, m-dimensions.

$$
\begin{aligned}
\mathbf{E}\Big[\sum_{j=1}^{m}(X_j - X_j')^2\Big] &= \mathbf{E}\Big[\sum_{j=1}^{m}\big({X_j}^2 - 2X_jX_j' + {X_j'}^2\big)\Big] \\
&= \mathbf{E}\Big[\sum_{j=1}^{m}{X_j}^2\Big] - 2\mathbf{E}\Big[\sum_{j=1}^{m}X_jX_j'\Big] + \mathbf{E}\Big[\sum_{j=1}^{m}{X_j'}^2\Big] \\
&= \sum_{j=1}^{m}\mathbf{E}\big[{X_j}^2\big] - 2\sum_{j=1}^{m}\mathbf{E}\big[X_jX_j'\big] + \sum_{j=1}^{m}\mathbf{E}\big[{X_j'}^2\big] \\
&= \sum_{j=1}^{m}\mathbf{E}\big[{X_j}^2\big] - 2\sum_{j=1}^{m}\mathbf{E}\big[X_j\big]\mathbf{E}\big[X_j'\big] + \sum_{j=1}^{m}\mathbf{E}\big[{X_j'}^2\big] \\
&= \sum_{j=1}^{m}({\mu_{1j}}^2 + \sigma^2) - 2\sum_{j=1}^{m}\mu_{1j}\mu_{2j} + \sum_{j=1}^{m}({\mu_{2j}}^2 + \sigma^2) \\
&= \sum_{j=1}^{m}{\mu_{1j}}^2 - 2\sum_{j=1}^{m}\mu_{1j}\mu_{2j} + \sum_{j=1}^{m}{\mu_{2j}}^2 + 2m\sigma^2
\end{aligned}
$$

5. The ratio of expected intra-class distance to inter-class distance is:

$$
ratio = \frac{2m\sigma^2}{\sum_{j=1}^{m}{\mu_{1j}}^2 - 2\sum_{j=1}^{m}\mu_{1j}\mu_{2j} + \sum_{j=1}^{m}{\mu_{2j}}^2 + 2m\sigma^2}
$$

The denominator can be re-written as follows:

$$
\begin{aligned}
denominator &= {\mu_{11}}^2 + \sum_{j=2}^{m}{\mu_{1j}}^2 - 2\Big(\mu_{11}\mu_{21} + \sum_{j=2}^{m}\mu_{1j}\mu_{1j}\Big) + {\mu_{21}}^2 + \sum_{j=2}^{m}{\mu_{1j}}^2 + 2m\sigma^2 \\
&= {\mu_{11}}^2 - 2\mu_{11}\mu_{21} + {\mu_{21}}^2 + 2m\sigma^2 \\
&= (\mu_{11} - \mu_{21})^2 + 2m\sigma^2
\end{aligned}
$$

$$\therefore ratio = \frac{2m\sigma^2}{(\mu_{11} - \mu_{21})^2 + 2m\sigma^2}$$

As m increases towards ∞, this ratio approaches 1. This indicates that the extra dimensions do not provide valuable information to help classify y as their number approaches infinity.

5 Fitting distributions with KL divergence

KL divergence for Gaussians.

1. The KL divergence between two univariate Gaussians is given as follows:

$$\begin{aligned}
KL(p(x)||q(x)) &= \mathbf{E}_p\left[\log \frac{p(x)}{q(x)}\right] \\
&= \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \\
&= \int_{-\infty}^{\infty} p(x) \log p(x) dx - \int_{-\infty}^{\infty} p(x) \log q(x) dx \\
&= \int_{-\infty}^{\infty} p(x) \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}(x-\mu_1)^2\right] dx \\
&- \int_{-\infty}^{\infty} p(x) \left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(x-\mu_2)^2\right] dx \\
&= \int_{-\infty}^{\infty} p(x) \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) dx - \int_{-\infty}^{\infty} p(x) \log\left(\frac{1}{\sqrt{2\pi}}\right) dx \\
&+ \int_{-\infty}^{\infty} p(x) \frac{(x-\mu_2)^2}{2} dx - \int_{-\infty}^{\infty} p(x) \frac{(x-\mu_1)^2}{2\sigma^2} dx \\
&= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \log\left(\frac{1}{\sqrt{2\pi}}\right) + \int_{-\infty}^{\infty} p(x) \left[\frac{(x-\mu_2)^2}{2} - \frac{(x-\mu_1)^2}{2\sigma^2}\right] dx \\
&= \mathbf{E}_p\left[\frac{(x-\mu_2)^2}{2} - \frac{(x-\mu_1)^2}{2\sigma^2}\right] + \log\left(\frac{1}{\sigma}\right) \\
&= \mathbf{E}_p[f(x, \mu_1, \mu_2, \sigma)] + g(\sigma) \\
&= \frac{1}{2}\mathbf{E}_p[(x-\mu_2)^2] - \frac{1}{2\sigma^2}\mathbf{E}_p[(x-\mu_1)^2] + \log\left(\frac{1}{\sigma}\right) \\
&= \frac{\mathbf{E}_p[x^2 - 2\mu_2 x + \mu_2^2]}{2} - \frac{1}{2} + \log\left(\frac{1}{\sigma}\right) \\
&= \frac{\mathbf{E}_p[x^2]}{2} - \mu_2 \mathbf{E}_p[x] + \frac{\mu_2^2}{2} - \frac{1}{2} + \log\left(\frac{1}{\sigma}\right) \\
&= \frac{\sigma^2 + \mu_1^2}{2} - \mu_1\mu_2 + \frac{\mu_2^2}{2} - \frac{1}{2} + \log\left(\frac{1}{\sigma}\right)
\end{aligned}$$

2. The value $\mu_1 = \mu_2$ minimizes $KL(p(x)||q(x))$. This follows from taking the partial derivative with respect to μ_1 of the above result for the KL divergence and setting it to 0.

$$\begin{aligned}
0 &= \frac{\partial KL(p(x)||q(x))}{\partial \mu_1} \\
0 &= \mu_1 - \mu_2 \\
\mu_1 &= \mu_2
\end{aligned}$$

6 Appendix

Matlab Code for 2.2

```
clc; clear; close all;

x = 0:0.001:6;
y = 6:-0.001:0;
[X,Y] = meshgrid(x,y);
D_x_1 = [1,4];
D_x_2 = [5,1];
D_o = [4,4];

p1 = sqrt((abs(X - D_x_1(1))).^2 + (abs(Y - D_x_1(2))).^2);
p2 = sqrt((abs(X - D_x_2(1))).^2 + (abs(Y - D_x_2(2))).^2);
p3 = sqrt((abs(X - D_o(1))).^2 + (abs(Y - D_o(2))).^2);
data(:,:,1) = p1;
data(:,:,2) = p2;
data(:,:,3) = p3;
[~,I] = min(data,[],3);
I(I == 1) = 0;
I(I == 2) = 0;
I(I == 3) = 1;
figure;
imagesc(I);
title('L-2 Norm');
axis off

clc; clearvars -except X Y D_x_1 D_x_2 D_o;

p1 = ((abs(X - D_x_1(1))) + (abs(Y - D_x_1(2))));
p2 = ((abs(X - D_x_2(1))) + (abs(Y - D_x_2(2))));
p3 = ((abs(X - D_o(1))) + (abs(Y - D_o(2))));
data(:,:,1) = p1;
data(:,:,2) = p2;
data(:,:,3) = p3;
[~,I] = min(data,[],3);
I(I == 1) = 0;
I(I == 2) = 0;
I(I == 3) = 1;
figure;
imagesc(I);
title('L-1 Norm');
axis off

clc; clearvars -except X Y D_x_1 D_x_2 D_o;

p1_1 = (abs(X - D_x_1(1)));
p1_2 = (abs(Y - D_x_1(2)));
p1 = max(p1_1, p1_2);
p2_1 = (abs(X - D_x_2(1)));
p2_2 = (abs(Y - D_x_2(2)));
p2 = max(p2_1, p2_2);
p3_1 = (abs(X - D_o(1)));
p3_2 = (abs(Y - D_o(2)));
```

```
p3 = max(p3_1, p3_2);
data(:,:,1) = p1;
data(:,:,2) = p2;
data(:,:,3) = p3;
[~,I] = min(data,[],3);
I(I == 1) = 0;
I(I == 2) = 0;
I(I == 3) = 1;
figure;
imagesc(I);
title('L-inf Norm');
axis off

clc; clearvars -except X Y D_x_1 D_x_2 D_o;

p1_temp(:,:,1) = (abs(X - D_x_1(1)));
p1_temp(:,:,2) = (abs(Y - D_x_1(2)));
p1 = sum(p1_temp ~= 0, 3);

p2_temp(:,:,1) = (abs(X - D_x_2(1)));
p2_temp(:,:,2) = (abs(Y - D_x_2(2)));
p2 = sum(p2_temp ~= 0, 3);

p3_temp(:,:,1) = (abs(X - D_o(1)));
p3_temp(:,:,2) = (abs(Y - D_o(2)));
p3 = sum(p3_temp ~= 0, 3);

data(:,:,1) = p1;
data(:,:,2) = p2;
data(:,:,3) = p3;
[~,I] = min(data,[],3);
I(I == 1) = 0;
I(I == 2) = 0;
I(I == 3) = 1;
figure;
imagesc(I);
title('L-0 Pseudo-Norm');
axis off
```