CIS 520, Machine Learning, Fall 2018: Assignment 1

Wentao He

September 16, 2018

Collaborators: N/A

1 Conditional independence in probability models

1. We can write $p(x_i)$ as:

$$\begin{aligned} p(x_i) &= \sum_{j=1}^{k} p(x_i \mid z_i = j)\pi_j \\ &= \sum_{j=1}^{k} f_j(x_i)\pi_j \end{aligned}$$

2. The formula for $p(x_1, \dots, x_n)$ is:

$$\begin{aligned} p(x_1, \dots, x_n) &= \sum_{z_1,\dots,z_n} p(x_1, \dots, x_n \mid z_1, \dots, z_n) \\ &= \sum_{z_1,\dots,z_n} p(x_1 \mid x_2, \dots, x_n, z_1, \dots, z_n) \dots p(x_n \mid z_1, \dots, z_n) p(z_1, \dots, z_n) \\ &= \sum_{z_1,\dots,z_n} \prod_{i=1}^{n} p(x_i \mid z_i) p(z_i) \\ &= \prod_{i=1}^{n} \sum_{j=1}^{k} p(x_1 \mid z_i \ = j\ )\pi_j \\ &= \prod_{i=1}^{n} \sum_{j=1}^{k} f_j(x_i)\pi_j \end{aligned}$$

3. The formula for $p(z_u = v \mid x_1, \dots, x_n)$ is:

$$\begin{aligned} p(z_u = v \mid x_1, \dots, x_n) &= \frac{p(x_1, \dots, x_n \mid z_u) p(z_u = v)}{p(x_1, \dots, x_n} \\ &= \frac{p(x_u \mid z_u = v)\pi_v \prod_{i=1, i\neq u}^{n} \sum_{j=1}^{k} p(x_i \mid z_i = j)\pi_j}{\prod_{i=1}^{n} \sum_{j=1}^{k} p(x_1 \mid z_i = j)\pi_j} \\ &= \frac{p(x_u \mid z_u = v)\pi_v}{\sum_{j=1}^{k} p(x_u \mid z_u = j)\pi_j} \\ &= \frac{f_v(x_u)\pi_v}{\sum_{j=1}^{k} f_j(x_u)\pi_j} \end{aligned}$$

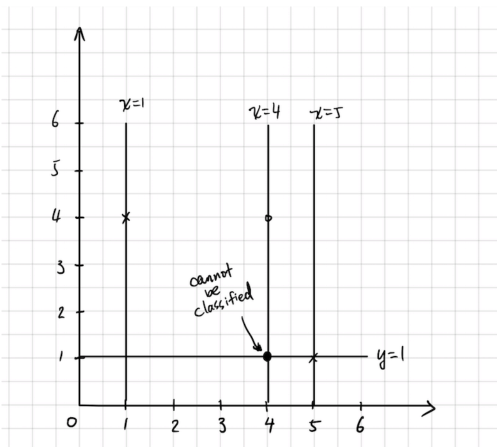
2 Non-Normal Norms

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

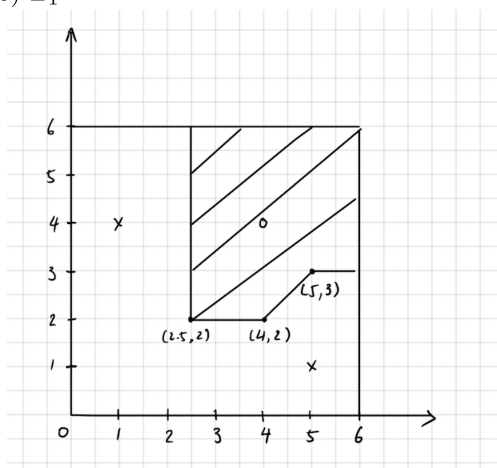
   a) L_0: x_4 with distance = 2
   b) L_1: x_3 with distance = 1.2
   c) L_2: x_2 with distance = 0.79
   d) L_{inf}: x_2 with distance = 0.6

2. Draw the 1-Nearest Neighbor decision boundaries with the given norms and lightly shade the o region:

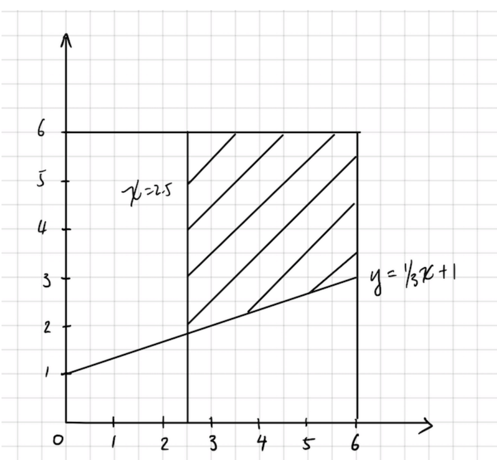
   a) L_0



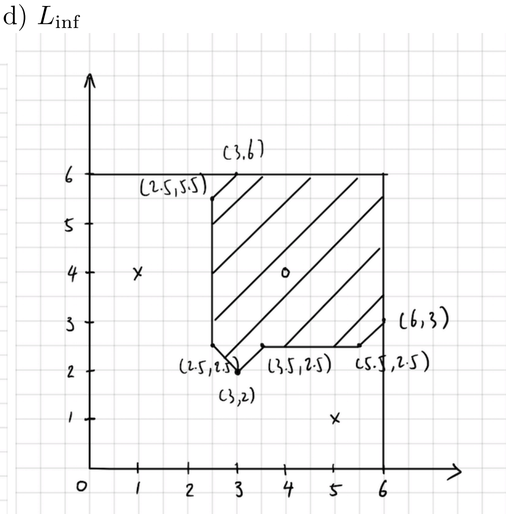
   b) L_1



   c) L_2



   d) L_{inf}



3 Decision trees

1. Concrete sample training data.

(a) The sample entropy $H(Y)$ is

$$\begin{aligned}
H(Y) &= -\sum_{i=1}^{k} P(Y = y_i) \log_2 P(Y = y_i) \\
&= -P(Y = +) \log_2 P(Y = +) - P(Y = -) \log_2 P(Y = -) \\
&= -(\frac{16}{30}) \log_2(\frac{16}{30}) - (\frac{14}{30}) \log_2(\frac{14}{30}) \\
&= 0.9968
\end{aligned}$$

(b) The information gains are $IG(X_1) = \underline{\mathbf{0.0114}}$ and $IG(X_2) = \underline{\mathbf{0.0487}}$.

To find the information gains $IG(X_1)$ and $IG(X_2)$, we must first find $H(Y \mid X_1)$ and $H(Y \mid X_2)$.

$H(Y \mid X_1)$ can be calculated as follows:

$$\begin{aligned}
H(Y \mid X_1) &= \sum_x P(X_1 = x) H(Y \mid X_1 = x) \\
&= -\sum_x P(X_1 = x) \sum_y P(Y = y \mid X_1 = x) \log_2 P(Y = y \mid X_1 = x) \\
&= -\sum_{x,y} P(X_1 = x, Y = y) \log_2 P(Y = y \mid X_1 = x) \\
&= -P(X_1 = T, Y = +) \log_2 P(Y = + \mid X_1 = T) - P(X_1 = T, Y = -) \log_2 P(Y = - \mid X_1 = T) \\
&\quad - P(X_1 = F, Y = +) \log_2 P(Y = + \mid X_1 = F) - P(X_1 = F, Y = -) \log_2 P(Y = - \mid X_1 = F)
\end{aligned}$$

The aforementioned probabilities can be calculated as follows:

$$\begin{aligned}
P(X_1 = T, Y = +) &= \frac{6}{30} \\
P(Y = + \mid X_1 = T) &= \frac{6}{13} \\
P(X_1 = T, Y = -) &= \frac{7}{30} \\
P(Y = - \mid X_1 = T) &= \frac{7}{13} \\
P(X_1 = F, Y = +) &= \frac{10}{30} \\
P(Y = + \mid X_1 = F) &= \frac{10}{17} \\
P(X_1 = F, Y = -) &= \frac{7}{30} \\
P(Y = - \mid X_1 = F) &= \frac{7}{17}
\end{aligned}$$

$H(Y \mid X_1) =$ can be calculated as follows:

$$\begin{aligned}
H(Y \mid X_1) &= -\frac{6}{30} \log_2 \frac{6}{13} - \frac{7}{30} \log_2 \frac{7}{13} - \frac{10}{30} \log_2 \frac{10}{17} - \frac{7}{30} \log_2 \frac{7}{17} \\
&= 0.9854
\end{aligned}$$

Therefore:

$$\begin{aligned} IG(X_1) &= H(Y) - H(Y \mid X_1) \\ &= 0.9968 - 0.9854 \\ &= 0.0114 \end{aligned}$$

Similarly, $H(Y \mid X_2)$ can be calculated as follows:

$$\begin{aligned} H(Y \mid X_2) = &\sum_x P(X_2 = x) H(Y \mid X_2 = x) \\ = &-\sum_x P(X_2 = x) \sum_y P(Y = y \mid X_2 = x) \log_2 P(Y = y \mid X_2 = x) \\ = &-\sum_{x,y} P(X_2 = x, Y = y) \log_2 P(Y = y \mid X_2 = x) \\ = &-P(X_2 = T, Y = +) \log_2 P(Y = + \mid X_2 = T) - P(X_2 = T, Y = -) \log_2 P(Y = - \mid X_2 = T) \\ &-P(X_2 = F, Y = +) \log_2 P(Y = + \mid X_2 = F) - P(X_2 = F, Y = -) \log_2 P(Y = - \mid X_2 = F) \end{aligned}$$

The aforementioned probabilities can be calculated as follows:

$$\begin{aligned} P(X_2 = T, Y = +) &= \frac{4}{30} \\ P(Y = + \mid X_2 = T) &= \frac{4}{11} \\ P(X_2 = T, Y = -) &= \frac{7}{30} \\ P(Y = - \mid X_1 = T) &= \frac{7}{11} \\ P(X_2 = F, Y = +) &= \frac{12}{30} \\ P(Y = + \mid X_2 = F) &= \frac{12}{19} \\ P(X_2 = F, Y = -) &= \frac{7}{30} \\ P(Y = - \mid X_2 = F) &= \frac{7}{19} \end{aligned}$$

$H(Y \mid X_2) =$ can be calculated as follows:

$$\begin{aligned} H(Y \mid X_2) = &-\frac{4}{30} \log_2 \frac{4}{11} - \frac{7}{30} \log_2 \frac{7}{11} - \frac{12}{30} \log_2 \frac{12}{19} - \frac{7}{30} \log_2 \frac{7}{19} \\ = &\ 0.9481 \end{aligned}$$

Therefore:

$$\begin{aligned} IG(X_2) &= H(Y) - H(Y \mid X_2) \\ &= 0.9968 - 0.9481 \\ &= 0.0487 \end{aligned}$$

The information gains are:

$$IG(X_1) = \underline{\mathbf{0.0114}}$$
$$IG(X_2) = \underline{\mathbf{0.0487}}$$

(c) The decision tree that would be learned is shown in Figure 1. Since $IG(X_2) > IG(X_1)$, this specific set of training example will be based on feature X_2:

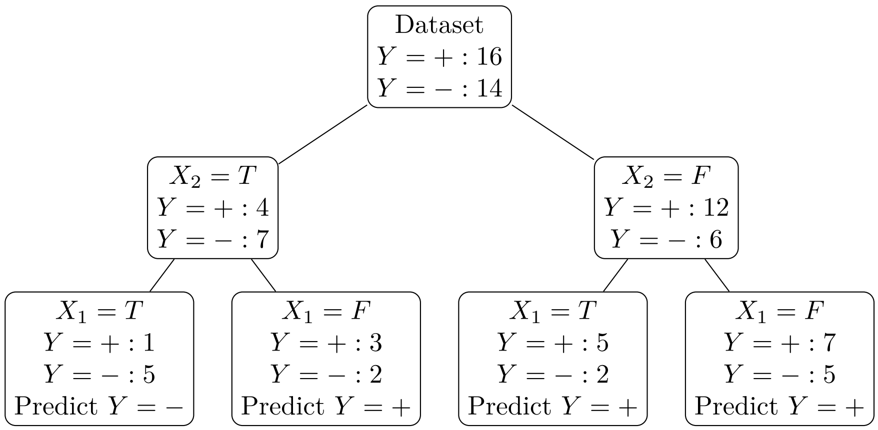


Figure 1: The decision tree that would be learned.

2. Information gain and KL-divergence.

(a) If variables X and Y are independent, is $IG(x, y) = 0$?

If X and Y are independent, $p(x, y) = P(x)P(y)$
which makes $\log \dfrac{(p(x)p(y)}{p(x, y)} = 0$.
Therefore $IG(x, y) = 0$.

(b) Proof that $IG(x, y) = H[x] - H[x \mid y] = H[y] - H[y \mid x]$.

Using conditional probability, $p(x, y) = p(x \mid y)p(y)$:

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

Using marginalization:

$$\begin{aligned}
IG(x,y) =&\ -\sum_x p(x,y)\log p(x) + \sum_x\sum_y p(x\mid y)p(y)\log p(x\mid y) \\
=&\ -\sum_x p(x,y)\log p(x) + \sum yp(y)\sum_x p(x\mid y)\log p(x\mid y) \\
=&\ -\sum_x p(x,y)\log p(x) + \sum_y p(Y=y)H(x\mid Y=y)
\end{aligned}$$

The first term of the above equation $-\sum_x p(x,y)\log p(x)$ is the negative of the sample entropy, $-H[x]$, and the negative of the second term of the above equation $-\sum_y p(Y=y)H(x\mid Y=y)$, which is the negative of the conditional entropy, $-H[x\mid y]$. Therefore, we have proved that this definition of information gain is equivalent to the one given in class:

$$IG(x,y) = H[x] - H[x\mid y]$$

4 High dimensional hi-jinx

(a) Intra-class distance.

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

(b) Inter-class distance.

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

(c) Intra-class distance, m-dimensions.

$$\begin{aligned}
\mathbf{E}[\sum_{j=1}^{m}(X_j - X_j')^2] &= E[(X_1 - X_1')^2 + (X_2 - X_2')^2 + (X_3 - X_3')^2 + \cdots + (X_m - X_m')^2] \\
&= (\mu_1 1^2 + \sigma^2 - 2\mu_1 1^2 + \mu_1 1^2 + \sigma^2) + (\mu_1 2^2 + \sigma^2 - 2\mu_1 2^2 + \mu_1 2^2 + \sigma^2) + \\
&\quad (\mu_1 3^2 + \sigma^2 - 2\mu_1 3^2 + \mu_1 3^2 + \sigma^2) + \cdots + (\mu_1 m^2 + \sigma^2 - 2\mu_1 m^2 + \mu_1 m^2 + \sigma^2) \\
&= 2m\sigma^2
\end{aligned}$$

(d) Inter-class distance, m-dimensions.

$$\begin{aligned}
\mathbf{E}[\sum_{j=1}^{m}(X_j - X_j')^2] &= E[(X_1 - X_1')^2 + (X_2 - X_2')^2 + (X_3 - X_3')^2 + \cdots + (X_m - X_m')^2] \\
&= (\mu_{11}^2 + \sigma^2 - 2\mu_{11}\mu_{21} + \mu_{2}1^2 + \sigma^2) + (\mu_{12}^2 + \sigma^2 - 2\mu_{12}\mu_{22} + \mu_{22}^2 + \sigma^2) + \\
&\quad (\mu_{13}^2 + \sigma^2 - 2\mu_{13}\mu_{23} + \mu_{23}^2 + \sigma^2) + \cdots + (\mu_{1m}^2 + \sigma^2 - 2\mu_{1m}\mu_{2m} + \mu_{2m}^2 + \sigma^2) \\
&= 2m\sigma^2 + \sum_{j=1}^{m}(\mu_{1j} - \mu_{2j})^2
\end{aligned}$$

(e) The ratio of expected intra-class distance to inter-class distance is: $\dfrac{2m\sigma^2}{(\mu_{11} - \mu_{21})^2 + 2m\sigma^2}$. As m increases towards ∞, this ratio approaches 1.

5 Fitting distributions with KL divergence

KL divergence for Gaussians.

1. The KL divergence between two univariate Gaussians is given by

$$\begin{aligned}
f &= \frac{(x-\mu_2)^2}{2} - \frac{(x-\mu_1)^2}{2\sigma^2} \\
g &= \log \frac{1}{\sigma}
\end{aligned}$$

The two univariate Gaussian distributions can be written as:

$$p(x) = \mathcal{N}(\mu_1, \sigma^2) = {\frac{1}{\sigma\sqrt{2\pi}}}^{\frac{-(x-\mu_1)^2}{2\sigma^2}}$$

$$q(x) = \mathcal{N}(\mu_2, 1) = {\frac{1}{\sqrt{2\pi}}}^{\frac{-(x-\mu_2)^2}{2}}$$

Therefore, the formula for $KL(p(x)||q(x))$ is:

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

2. The value $\underline{\boldsymbol{\mu_1 = \mu_2}}$ minimizes $KL(p(x)||q(x))$, which is $\underline{\boldsymbol{\frac{1}{2}\sigma^2 - \frac{1}{2} + \log\frac{1}{\sigma}}}$.

First, we know that:

$$KL(p(x)||q(x)) = E_p\left(\frac{(x-\mu_2)^2}{2} - \frac{(x-\mu_1)^2}{2\sigma^2}\right) + \log\frac{1}{\sigma}\Bigg)$$

Also, we know that:

$$E_p[x^2] = \mu_1^2 + \sigma^2$$
$$E_p[x] = \mu_1$$

Therefore, $KL(p(x)||q(x))$ becomes:

$$KL(p(x)||q(x)) = \frac{1}{2}\mu_1^2 + \frac{1}{2}\sigma^2 - \mu_1\mu_2 + \frac{1}{2}\mu_2^2 - \frac{1}{2} + \log\frac{1}{\sigma}$$

For a minimum value of $KL(p(x)||q(x))$, we set its derivative equal to zero:

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

And we get:

$$\mu_1 = \mu_2$$

When $\mu_1 = \mu_2$:

$$KL(p(x)||q(x)) = \frac{1}{2}\sigma^2 - \frac{1}{2} + \log\frac{1}{\sigma}$$