CIS 520: Problem Set #1

Due on September 15, 2017

Lyle Ungar

Eric Oh

Collaborators: Jiarui Lu

Problem 1

High Dimensional Hi-Jinx

Solution

1. (Intra-class distance)

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

2. (Inter-class distance)

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

3. (Intra-class distance, m dimensions)

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

4. (Inter-class distance, m dimensions)

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

5. Under the assumption that only one dimension is informative about the class values, the ratio of expected intra-class distanced divided by inter-class distance is given by

$$\frac{2m\sigma^2}{2m\sigma^2 + \sum_{j=1}^{m}(\mu_{1j} - \mu_{2j})^2} = \frac{2m\sigma^2}{2m\sigma^2 + (\mu_{11} - \mu_{21})^2}$$

where the equality follows from the assumption. Then as m gets large, we have

$$lim_{m\to\infty}\frac{2m\sigma^2}{2m\sigma^2 + (\mu_{11} - \mu_{21})^2} = lim_{m\to\infty}\frac{2\sigma^2}{2\sigma^2} = 1$$

where the first equality follows from L'Hopital's Rule.

Problem 2

Non-Normal Norms

Solution

1. These are the distances to x_1 under the following norms:

   - L_0

     $$\|x_2 - x_1\|_0 = 2 \qquad \|x_3 - x_1\|_0 = 4 \qquad \|x_4 - x_1\|_0 = 4$$

     Thus, x_2 is the closest to x_1 under the L_0 norm.

   - L_1

     $$\|x_2 - x_1\|_1 = 5.4 \qquad \|x_3 - x_1\|_1 = 3.0 \qquad \|x_4 - x_1\|_1 = 3.3$$

     Thus, x_3 is the closest to x_1 under the L_1 norm.

   - L_2

     $$\|x_2 - x_1\|_2 = 4.8 \qquad \|x_3 - x_1\|_2 = 1.8 \qquad \|x_4 - x_1\|_2 = 1.9$$

     Thus, x_3 is the closest to x_1 under the L_2 norm.

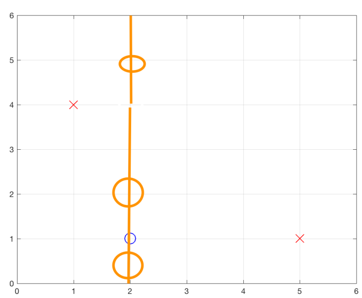
   - L_{inf}

     $$\|x_2 - x_1\|_\infty = 4.7 \qquad \|x_3 - x_1\|_\infty = 1.5 \qquad \|x_4 - x_1\|_\infty = 1.4$$

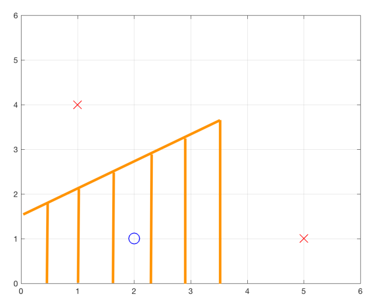
     Thus, x_4 is the closest to x_1 under the L_∞ norm.

2. Below are the 1-Nearest Neighbor decision boundaries for the given norms.

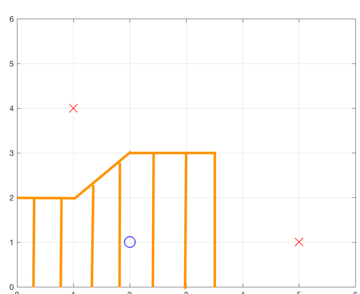
   - L_0

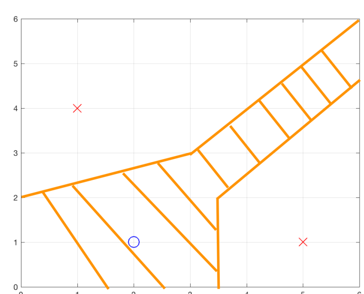
   - L_2

   - L_1

   - L_∞

Problem 3

Conditional Independence in Probability Models

Solution

1. The formula for $p(x_i)$ is given by

$$p(x_i) = \sum_j p(x_i, z_i = j) = \sum_j p(x_i | z_i = j) P(z_i = j) = \sum_j f_j(x_i)\pi_j$$

2. The formula for $p(x_1, \ldots, x_n)$ is given by

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \sum_j f_j(x_i)\pi_j$$

using the independence of $x_1, \ldots, x_n$.

3. The formula for $p(z_u = v | x_1, \ldots, x_n)$ is given by

$$p(z_u = v | x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n | z_u = v) p(z_u = v)}{p(x_1, \ldots, x_n)}$$

using Bayes rule. We know that $p(z_u = v) = \pi_v$. Part 2 of this question gives the formula for the denominator. WLOG, assume $1 < u < n$. Then the remaining term in the numerator is given by

$$\begin{aligned} p(x_1, \ldots, x_n | z_u = v) &= \prod_{i=1}^{n} p(x_i | z_u = v) = \prod_{i=1}^{u-1} p(x_i | z_u = v) \times p(x_u | z_u = v) \times \prod_{u+1}^{n} p(x_i | z_u = v) \\ &= \prod_{i=1}^{u-1} p(x_i) \times f_v(x_u) \times \prod_{u+1}^{n} p(x_i) \qquad \text{(By independence of } x_i\text{)} \\ &= \left(\prod_{i=1}^{u-1} \sum_j f_j(x_i)\pi_j \right) \times f_v(x_u) \times \left(\prod_{u+1}^{n} \sum_j f_j(x_i)\pi_j \right) \end{aligned}$$

Then it follows that

$$\begin{aligned} p(z_u = v | x_1, \ldots, x_n) &= \frac{\left(\prod_{i=1}^{u-1} \sum_j f_j(x_i)\pi_j\right) \times f_v(x_u) \times \left(\prod_{u+1}^{n} \sum_j f_j(x_i)\pi_j\right) \times \pi_v}{\prod_{i=1}^{n} \sum_j f_j(x_i)\pi_j} \\ &= \frac{f_v(x_u)\pi_v}{\sum_j f_j(x_u)\pi_j} \end{aligned}$$

Problem 4

Fitting distributions with KL divergence

Solution

1. The KL divergence between two univariate Gaussian distributions $p(x) \sim \mathrm{N}(\mu_1, \sigma^2)$ and $q(x) \sim \mathrm{N}(\mu_2, 1)$

is given by

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

where $f(x,\mu_1,\mu_2,\sigma) = -\frac{1}{2\sigma^2}(x-\mu_1)^2+\frac{1}{2}(x-\mu_2)^2$ and $g(\sigma)=\log(\sigma^{-1})$.

2. For fixed μ_2 and σ, the value of μ_1 that minimizes $KL(p(x)\|q(x))$ is found by taking the derivative of the expression in Part 1 w.r.t μ_1.

$$\begin{aligned}
KL(p(x)\|q(x)) &= \mathrm{E}_p\left[-\frac{1}{2\sigma^2}(x-\mu_1)^2+\frac{1}{2}(x-\mu_2)^2\right]+\log(\sigma^{-1}) = -\frac{1}{2\sigma^2}\mathrm{E}(x-\mu_1)^2+\frac{1}{2}\mathrm{E}(x-\mu_2)^2+\log(\sigma^{-1})\\
&= -\frac{\sigma^2}{2\sigma^2}+\frac{1}{2}\mathrm{E}\left(x-\mu_1+\mu_1-\mu_2\right)^2+\log(\sigma^{-1})\\
&= -\frac{1}{2}+\frac{1}{2}\mathrm{E}\left((x-\mu_1)^2+2(x-\mu_1)(\mu_1-\mu_2)+(\mu_1-\mu_2)^2\right)+\log(\sigma^{-1})\\
&= -\frac{1}{2}+\frac{1}{2}\left[\mathrm{E}(x-\mu_1)^2+2\mathrm{E}(x-\mu_1)(\mu_1-\mu_2)+\mathrm{E}(\mu_1-\mu_2)^2\right]+\log(\sigma^{-1})\\
&= -\frac{1}{2}+\frac{1}{2}\left[\sigma^2+(\mu_1-\mu_2)^2\right]+\log(\sigma^{-1}) \qquad \text{(Middle term above is 0)}
\end{aligned}$$

$$\frac{d}{d\mu_1}KL(p(x)\|q(x)) = \mu_1-\mu_2 \equiv 0 \quad \Rightarrow \quad \mu_1=\mu_2$$

Thus, setting μ_1 equal to μ_2 minimizes $KL(p(x)\|q(x))$ for fixed μ_2 and σ. At this minimum, the value is $-\frac{1}{2}+\frac{\sigma^2}{2}+\log(\sigma^{-1})$.

Problem 5

Decision Trees

Solution

1. The sample entropy $H(Y)$ for this training data is given by

$$H(Y) = -\left[P(Y=+)\log(P(Y=+))+P(Y=-)\log(P(Y=-))\right] = -\left[\frac{13}{25}\log\left(\frac{13}{25}\right)+\frac{12}{25}\log\left(\frac{12}{25}\right)\right] = 0.9988$$

2. The conditional entropies $H(Y|X_1)$ and $H(Y|X_2)$ are given by

$$\begin{aligned}
H(Y|X_1) &= P(X_1=T)H(Y|X_1=T)+P(X_1=F)H(Y|X_1=F)\\
&= \frac{11}{25}\left[-\left\{\frac{4}{11}\log\left(\frac{4}{11}\right)+\frac{7}{11}\log\left(\frac{7}{11}\right)\right\}\right]+\frac{14}{25}\left[-\left\{\frac{9}{14}\log\left(\frac{9}{14}\right)+\frac{5}{14}\log\left(\frac{5}{14}\right)\right\}\right]\\
&= 0.9427
\end{aligned}$$

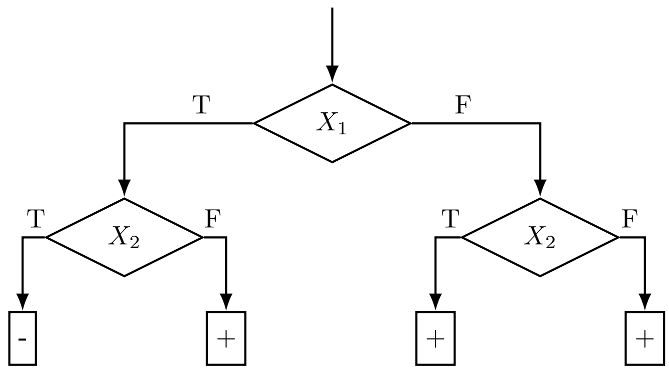
and

$$\begin{aligned}
H(Y|X_2) &= P(X_2 = T)H(Y|X_2 = T) + P(X_2 = F)H(Y|X_2 = F) \\
&= \frac{11}{25}\left[-\left\{\frac{5}{11}\log\left(\frac{5}{11}\right) + \frac{6}{11}\log\left(\frac{6}{11}\right)\right\}\right] + \frac{14}{25}\left[-\left\{\frac{8}{14}\log\left(\frac{8}{14}\right) + \frac{6}{14}\log\left(\frac{6}{14}\right)\right\}\right] \\
&= 0.9891
\end{aligned}$$

Thus the information gains are

$$IG(X_1) = H(Y) - H(Y|X_1) = .0561 \qquad IG(X_2) = H(Y) - H(Y|X_2) = .0097$$

3. The decision tree learned by ID3 is given by



4. If variables X and Y are independent, then $IG(x, y) = 0$. We prove this by using the definition of independence given by $p(x, y) = p(x)p(y)$ in the KL-divergence information gain.

$$\begin{aligned}
IG(x, y) \equiv KL(p(x)\|q(x)) &= -\sum_x \sum_y p(x, y)\log\left(\frac{p(x)p(y)}{p(x, y)}\right) \\
&= -\sum_x \sum_y p(x, y)\log\left(\frac{p(x)p(y)}{p(x)p(y)}\right) \\
&= -\sum_x \sum_y p(x, y) \times 0 \\
&= 0
\end{aligned}$$

5. We can show that the definition of information gain using the KL divergence is equivalent to the definition involving entropy, ie. $IG(x, y) = H(x) - H(x|y) = H(y) - H(y|x)$. We first show that $IG(x, y) = H(y) - H(y|x)$.

$$\begin{aligned}
IG(x, y) &= -\sum_x \sum_y p(x, y)\log\left(\frac{p(x)p(y)}{p(x, y)}\right) = -\sum_x \sum_y p(y|x)p(x)\log\left(\frac{p(x)p(y)}{p(y|x)p(x)}\right) \\
&= -\sum_x p(x)\sum_y p(y|x)\left\{\log(p(y)) - \log(p(y|x))\right\} \\
&= -\sum_x \sum_y p(x, y)\log(p(y)) + \sum_x p(x)\sum_y p(y|x)\log(p(y|x)) \\
&= -\sum_y \sum_x p(x, y)\log(p(y)) - \sum_x p(x)H(Y|X = x) \\
&= -\sum_y p(y)\log(p(y)) - \sum_x p(x)H(Y|X = x) \\
&= H(Y) - H(Y|X)
\end{aligned}$$

Note that we could have chosen to write $p(x,y) = p(x|y)p(y)$ in the first line of the proof, giving us

$$IG(x,y) = -\sum_x \sum_y p(x,y) \log\left(\frac{p(x)p(y)}{p(x,y)}\right) = -\sum_x \sum_y p(x|y)p(y) \log\left(\frac{p(x)p(y)}{p(x|y)p(y)}\right)$$

The proof then follows by symmetry and we have shown that $IG(x,y) = H(x) - H(x|y)$. To show that $H(x) - H(x|y) = H(y) - H(y|x)$, it suffices to show that

$$-\sum_x \sum_y p(y|x)p(x) \log\left(\frac{p(x)p(y)}{p(y|x)p(x)}\right) = -\sum_x \sum_y p(x|y)p(y) \log\left(\frac{p(x)p(y)}{p(x|y)p(y)}\right)$$

Bayes rule gives us $p(y|x)p(x) = p(x|y)p(y)$, so the above equality follows immediately.

Note, the above proof can also be done using continuous x and y. In this case, the sums are replaced by integrals and we would need to invoke Fubini's Theorem to change the order of integration. Checking the conditions for Fubini's Theorem is beyond the scope of this class; thus, the rest of the proof would follow as is.