CIS 520, Machine Learning, Fall 2011: Assignment 1

Zheyin Zhao

September 22, 2011

1 High dimensional hi-jinx

1. Intra-class distance.

$$E[(X-X')^2] = E[(X^2 + X'^2 - 2XX'] = E[X^2] + E[X'^2] - 2E[X]E[X']$$
$$= \mu_1^2 + \sigma^2 + \mu_1^2 + \sigma^2 - 2\mu_1^2 = 2\sigma^2$$

2. Inter-class distance.

$$E[(X-X')^2] = E[(X^2 + X'^2 - 2XX'] = E[X^2] + E[X'^2] - 2E[X]E[X']$$
$$= \mu_1^2 + \sigma^2 + \mu_2^2 + \sigma^2 - 2\mu_1\mu_2$$

3. Intra-class distance, m-dimensions.

$$\mathbf{E}[\sum_{j=1}^{m}(X_j - X_j')^2] = \sum_{j=1}^{m} E[(X_j - X_j^{'})^2]$$
$$= 2m\sigma^2$$

4. Inter-class distance, m-dimensions.

$$\mathbf{E}[\sum_{j=1}^{m}(X_j - X_j')^2] = \sum_{j=1}^{m} E[(X_j - X_j^{'})^2]$$
$$= \sum_{j=1}^{m}(\mu_{1j} - \mu_{2j})^2 + 2m\sigma^2$$

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

2 Fitting distributions with KL divergence

1. KL divergence for Gaussians.

   (a) The KL divergence between two univariate Gaussians is given by $f = 1/2(x-\mu_2)^2 - 1/2\sigma^2(x-\mu_2)^2$ and $g = -\log\sigma$.

$$KL(p(x)||q(x)) = \int_{-\infty}^{\infty} N(\mu_1,\sigma^2)\log\frac{N(\mu_1,\sigma^2)}{N(\mu_2,1)}dx$$
$$= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu_1)^2}{2\sigma}}[1/2(x-\mu_2)^2 - 1/2\sigma^2(x-\mu_2)^2]dx + -\log\sigma$$
$$= \mathbf{E}_p[f(x,\mu_1,\mu_2,\sigma)] + g(\sigma)$$

(b) The value $\mu_1 = \mu_2$ minimizes $KL(p(x)||q(x))$.

$$
\begin{aligned}
0 =& \frac{\partial KL(p(x)||q(x))}{\partial \mu_1} \\
0 =& \frac{\partial E[f(x,\mu_1,\mu_2,\sigma)]}{\partial \mu_1} + 0 = \frac{\partial}{\partial \mu_1} E[\frac{1}{2}x^2 + \frac{1}{2}\mu_2^2 - x\mu_2 - \frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}\mu_1^2 + \frac{1}{\sigma^2}x\mu_1] \\
=& \frac{\partial}{\partial \mu_1}[\frac{1}{2}(\mu_1^2+\sigma^2) + \frac{1}{2}\mu_2^2 - \mu_1\mu_2 - \frac{1}{2\sigma^2}(\mu_1^2+\sigma^2) - \frac{1}{2\sigma^2}\mu_1^2 + \frac{1}{\sigma^2}\mu_1^2] = \mu_1 - \mu_2 \\
\mu_1 =& \mu_2
\end{aligned}
$$

2. KL divergence for Multinomials.

   (a) The KL divergence between two Multinomials is: $KL(p(x)||q(x)) = \sum_{i\ odd} \beta log\frac{\beta}{\theta_i} + \sum_{i\ even} \alpha log\frac{\alpha}{\theta_i}$.

   (b) The values $\alpha = \frac{(\prod_{i=even}\theta_i/\prod_{i=odd}\theta_i)^{1/n}}{n((\prod_{i=even}\theta_i/\prod_{i=odd}\theta_i)^{1/n}+1)}$ and $\beta = \frac{1}{n((\prod_{i=even}\theta_i/\prod_{i=odd}\theta_i)^{1/n}+1)}$ minimize $KL(p(x)||q(x))$.

$$
\begin{aligned}
\text{Lagrangian}\ \ \mathcal{L} &= KL(p(x)||q(x)) + \lambda(n(\alpha+\beta)-1) \\
&= \sum_{i=even} \alpha \log\frac{\alpha}{\theta_i} + \sum_{i=odd} \beta \log\frac{\beta}{\theta_i} + \lambda(n(\alpha+\beta)-1) \\
\frac{\partial}{\partial \alpha}\mathcal{L} &= \sum_{i=odd} \log\frac{\beta}{\theta_i} + n + \lambda n = 0 \\
\frac{\partial}{\partial \beta}\mathcal{L} &= \sum_{i=even} \log\frac{\alpha}{\theta_i} + n + \lambda n = 0 \\
\frac{\partial}{\partial \lambda}\mathcal{L} &= n(\alpha+\beta) - 1 = 0
\end{aligned}
$$

3 Conditional independence in probability models

1. We can write $p(x_i) = \sum_{j=1}^k p(x_i|z_i=j)p(z_i=j) = \sum_{j=1}^k f_j(x_i)\pi_j$ because the probablity of x take the value of x_{i} equals the sum of probablity of x=x_{i} in every possible distribution, which is marginalization.

2. The formula for $p(x_1,\ldots,x_n)$ is $\prod_{i=1}^n \sum_{j=1}^k f_j(x_i)\pi_j$ by the derivation below (simply by applying the chain rule).

$$
\begin{aligned}
p(x_1,\ldots,x_n) &= \prod_{i=1}^n p(x_i) \\
&= \prod_{i=1}^n \sum_{j=1}^k f_j(x_i)\pi_j
\end{aligned}
$$

3. The formula for $p(z_u = v \mid x_1,\ldots,x_n)$ is $\frac{f_v(x_u)\pi_v}{\sum_{j=1}^k f_v(x_u)\pi_j}$ by the derivation below (simply by applying the bayes rule).

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

4 Decision trees

1. Concrete sample training data.

   (a) The sample entropy $H(Y)$ is 0.985.

$$\begin{aligned} H(Y) &= -p(y=+)\log(p=+) - p(y=-)\log p(y=-) \\ &= -4/7\log 4/7 - 3/7\log 3/7 \\ &= 0.985 \end{aligned}$$

   (b) The information gains are $IG(X_1) = 0.183$ and $IG(X_2) = 0.045$.

$$\begin{aligned} IG(X_1) &= H(y) - H(y|x_1) \\ &= H(y) - (-7/21\log 7/8 - 1/21\log 1/8 - 5/21\log 5/13 - 8/21\log 8/13) \\ &= 0.985 - 0.802 = 0.183 \\ IG(X_2) &= H(y) - H(y|x_2) \\ &= H(y) - (-7/21\log 7/10 - 3/21\log 3/10 - 5/21\log 5/11 - 6/21\log 6/11) \\ &= 0.985 - 0.940 = 0.045 \end{aligned}$$

   (c) The decision tree that would be learned is shown in Figure 1.

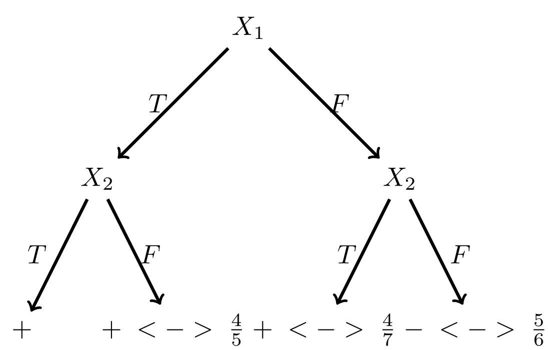

Figure 1: The decision tree that would be learned.

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:

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