CIS 520, Machine Learning, Fall 2011: Assignment 1

Your name here

September 16, 2011

1 High dimensional hi-jinx

1. Intra-class distance.

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

2. Inter-class distance.

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

3. Intra-class distance, m-dimensions.

$$\begin{aligned}\mathbf{E}[\sum_{j=1}^{m}(X_j - X'_j)^2] &= \dots \\ &= \dots\end{aligned}$$

4. Inter-class distance, m-dimensions.

$$\begin{aligned}\mathbf{E}[\sum_{j=1}^{m}(X_j - X'_j)^2] &= \dots \\ &= \dots\end{aligned}$$

5. The ratio of expected intra-class distance to inter-class distance is: .... As m increases towards ∞, this ratio approaches ....

2 Fitting distributions with KL divergence

1. KL divergence for Gaussians.

   (a) The KL divergence between two univariate Gaussians is given by $f = \dots$ and $g = \dots$.

$$\begin{aligned}KL(p(x)||q(x)) &= \dots \\ &= \dots \\ &= \mathbf{E}_p[f(x, \mu_1, \mu_2, \sigma)] + g(\sigma)\end{aligned}$$

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

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

2. KL divergence for Multinomials.
   (a) The KL divergence between two Multinomials is: $KL(p(x)||q(x)) = \ldots$.
   (b) The values $\alpha = \ldots$ and $\beta = \ldots$ minimize $KL(p(x)||q(x))$.

$$\begin{aligned} \text{Lagrangian } \mathcal{L} &= \ldots \\ &= \ldots \end{aligned}$$

3 Conditional independence in probability models

1. We can write $p(x_i) = \ldots$ because ….
2. The formula for $p(x_1, \ldots, x_n)$ is … by the derivation below.

$$\begin{aligned} p(x_1, \ldots, x_n) &= \ldots \\ &= \ldots \end{aligned}$$

3. The formula for $p(z_u = v \mid x_1, \ldots, x_n)$ is … by the derivation below.

$$\begin{aligned} p(z_u = v \mid x_1, \ldots, x_n) &= \ldots \\ &= \ldots \end{aligned}$$

4 Decision trees

1. Concrete sample training data.
   (a) The sample entropy $H(Y)$ is ….

$$\begin{aligned} H(Y) &= \ldots \\ &= \ldots \\ &= \ldots \end{aligned}$$

   (b) The information gains are $IG(X_1) = \ldots$ and $IG(X_2) = \ldots$.

$$\begin{aligned} IG(X_1) &= \ldots \\ IG(X_2) &= \ldots \end{aligned}$$

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

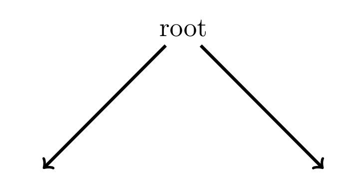

Figure 1: The decision tree that would be learned.

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:

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