10702/36702 Statistical Machine Learning, Spring 2008: Homework 5 Solutions

May 9, 2008

1 [20 points], (Robin)

★ **SOLUTION:**

(a) Take any set of N atoms J_N from the total set $\mathcal{D}$. The distance between f and g is then:

$$\begin{aligned}
||f-g|| &= ||f-f_N|| \\
&= ||\sum_{\mathcal{D}} \beta_j \psi_j - \sum_{J_N} \beta_j \psi_j|| \\
&= ||\sum_{\mathcal{D}-J_N} \beta_j \psi_j|| \\
&= \sqrt{\sum_{\mathcal{D}-J_N} \beta_j^2 < \psi_j, \psi_j >} \\
&= \sqrt{\sum_{\mathcal{D}-J_N} \beta_j^2}
\end{aligned}$$

which is minimized when J_N is over the N largest values of $|\beta_j|$.

(b) In OGA, at each step $r_{N-1} = f - f_N = \sum_{\mathcal{D}-J_{N-1}} \beta_j \psi_j$ where J_{N-1} is a set of functions selected so far. To choose the next function, compute

$$\begin{aligned}
cp_{N,i} &= | < r_{N-1}, \psi_i > | \\
&= | \sum_{\mathcal{D}-J_{N-1}} \beta_j < \psi_j, \psi_i > | \\
&= |\beta_i < \psi_i, \psi_i > + \sum_{\mathcal{D}-J_{N-1}-i} \beta_j < \psi_j, \psi_i > | \\
&= |\beta_i|
\end{aligned}$$

The maximum is achieved at $argmax_i |b_i|$. Hence OGA recovers f_N exactly.

(c)

$$\begin{aligned}
\sigma_N(f) &= ||f - f_N|| = \sqrt{\sum_{\mathcal{D}-J_N} \beta_j^2} \\
&< \sqrt{\sum_{N+1}^{|\mathcal{D}|} \frac{C^2}{j^{2/p}}} < \sqrt{\sum_{N+1}^{|\mathcal{D}|} \frac{C^2}{(N+1)^{2/p}}} \\
&< \sqrt{\frac{|\mathcal{D}|C^2}{(N+1)^{2/p}}} < \sqrt{\frac{(N+1)C^2}{(N+1)^{2/p}}} \\
&= \sqrt{\frac{C^2}{(N+1)^{\frac{2}{p-1}}}} = O(\frac{1}{N^{1/p-1/2}}) \\
&= O(\frac{1}{N^s})
\end{aligned}$$

2 [20 points], (Robin)

★ **SOLUTION:** Start with Bernstein's inequality

$$P(|\bar{X}_n| > t) \leq 2e^{-\frac{nt^2}{2\sigma^2 + \frac{2ct}{3}}}$$

Then substitute $t = \sigma\sqrt{\frac{2\delta}{n}} + \frac{2c\delta}{3n}$ and simplify to get

$$P(|\bar{X}_n| > \sigma\sqrt{\frac{2\delta}{n}} + \frac{2c\delta}{3n}) \leq 2e^{-\delta}$$

3 [20 points], (Robin)

★ **SOLUTION:**

(a) When $r = 1$ we have $u(x) - l(x) \leq \epsilon$, i.e. the bracket is in the ball of $\epsilon/2$ around $(u+l)/2$, the center of the bracket. This is also the $\epsilon/2$ cover. Since the bracket is contained in within this ball, we need more ϵ-brackets than the $\epsilon/2$ cover to cover the function. Hence,

$$N_1(\epsilon/2, \mathcal{F}) \leq N_{[]}(\epsilon, \mathcal{F}, L_1(P))$$

From Theorem 1.46 in notes we have

$$\begin{aligned}
P(sup_{f\in\mathcal{F}}|P_n(f) - P(f)| > \epsilon) &\leq 8N_1(\epsilon/8, \mathcal{F})e^{-n\epsilon^2/128B^2} \\
&\leq 8N_{[]}(\epsilon/4, \mathcal{F}, L_1(P))e^{-n\epsilon^2/32B^2} \\
&\to 0 \qquad \text{as} \quad n \to \infty
\end{aligned}$$

Hence proved.

(b) we can create the ϵ-brackets as follows. Let $-\infty = t_0 < t_1 < ... < t_k = \infty$ where $t \in R$. We can choose the bracketing functions themselves to be indicator functions. Let $I_{t_i} = I_{(-\infty, t_i]}$. For ϵ-bracket, we have $\int_{-\infty}^{\infty}(u(x) - l(x))p(x)dx \leq \epsilon$. Hence,

$$\int_{-\infty}^{\infty}(I_{t_i}(x) - I_{t_{i-1}}(x))p(x)dx \leq \epsilon \qquad \forall i = 0, 1, .., k$$

But if $x < t_{i-1}$ then $I_{t_i}(x) = I_{t_{i-1}}(x)$ and if $x \geq t_{i-1}$ then $I_{t_i}(x) - I_{t_{i-1}}(x) = 1$. So, $\int_{t_{i-1}}^{t_i} p(x) \leq \epsilon$. Hence, each bracket consumes a probability mass of atmost ϵ and since the total probability mass is 1, there are $1/\epsilon$ such brackets consuming the entire mass of 1 which is bounded by $2/\epsilon$ (if $1/\epsilon$ is not a whole number). Hence proved.

4 [20 points], (Robin)

★ **SOLUTION:** The VC dimension of axis-aligned rectangles in R^d is $2d$.

(1) Show that the $VC - dim \geq 2d$.

Consider a set of $2d$ points where each point only has one of the d dimensions set to either 1 or -1 and 0 for all other dimensions. It is easy to see that any subset of these points can be shattered by an axis-aligned rectangle. Hence the VC-dim is atleast 2d.

(2) Show that the $VC - dim < 2d + 1$. Consider a set of 2d+1 points. Consider finding the minimum and maximum of value in each dimension for these set of points and then building a R^d rectangle with these bounds. Since there are $2d + 1$ points, atleast one point must lie inside this rectangle. If we label this interior point as negative then there is no rectangle that can separate this labeling. This proves that $VC - dim < 2d + 1$.

Combining (1) and (2) we get that the $VC - dim = 2d$. Intuitively, there are $2d$ free parameters (lower and upper bound in each dimension) of the rectangle and hence the VC-dimension is $2d$.

5 [20 points], (Robin)

★ **SOLUTION:**

(a) You can assume k=6 since we need to compare the predicted and true clusters. Also, you need to define a good comparison metric.

One such metric is to penalize a pair of points that belongs to the same true cluster but is present in different predicted clusters and vice versa. You can then report the probability of error over all pairs of points in the data.

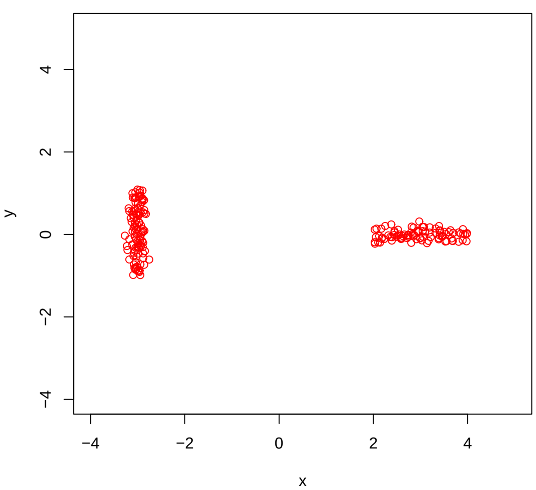
Under this scoring metric, K-Means error is 0.13 and Hierarchical clustering error is 0.16. This shows that K-Means performs better on the data.

(b) Note that here we need to select a subset of original features and not use any kind of projection or transformation of data. Some possibilities for feature selection are: (1) apply kmeans using each feature and select the set of features greedily with minimum distortion in predicted clusters (2) assume that the clusters must have similar number of points ad hence use entropy measure to select the features.

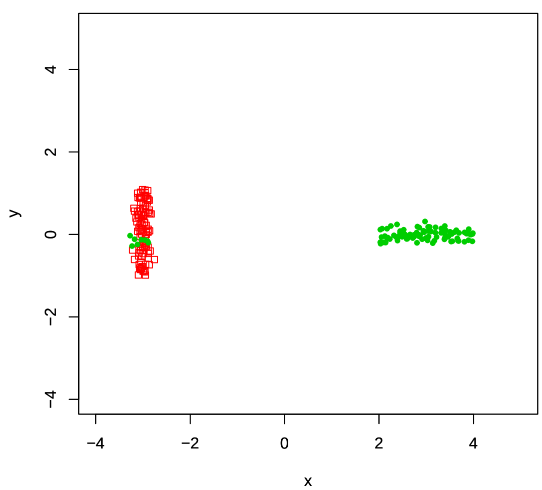
Using the first technique, K-Means error with top 15 features is 0.15 which is only 2% more than using all features, i.e. with a only quarter of original features.

6 [20 points], (Robin)

★ **SOLUTION:**



(a) Original data



(b) Clustered data