CSE 417T Machine Learning

Hw6, Due: Nov.22, 2016

Chih Yun Pai

1.

(a) Show that maximizing the divergence of the label distribution in the child nodes to the uniform distribution is equivalent to minimizing entropy.

Sol:

$\text{P: true distribution data}$

$\text{Q: model, approximation of P}$

$\text{Entropy(S)} = -\sum_{i=1}^{k} p_i \, log_2 p_i$

$D_{KL}(P||Q) = \sum_{i=1}^{k} p_i \, log_2 \frac{p_i}{Q} = \sum_{i=1}^{k} p_i \, log_2 p_i - \sum_{i=1}^{k} p_i \, log_2 Q \ \ (contant)$

$= \sum_{i=1}^{k} p_i \, log_2 p_i - log_2 Q \ \sum_{i=1}^{k} p_i$

$\Rightarrow \min Entropy \sim \max D_{KL}(P||Q)$

(b) Show that maximizing information gain is equivalent minimizing entropy.

S splits to S_L and S_R

$min_S \ H(S_L, S_R) = min_S \left[\frac{|S_L|}{|S|} \mathrm{H}(S_L) + \frac{|S_R|}{|S|} \mathrm{H}(S_R)\right]$

$= max_A \ G(S, A) = max_A\left[H(S) - \sum_{j=1}^{k\alpha} P(A = j) \, \mathrm{H}(S_j)\right]$

$= max_A\left[H(S) - P(A = i)H(S_j) - P(A \neq j)H(S_j)\right]$

$= max_A \left[H(S) - \frac{|S_A|}{|S|} H(S_j) - \frac{|S_{A'}|}{|S|} H(S_j)\right]$

$= \mathrm{H(S)} - min_A \frac{|S_A|}{|S|} H(S_j)$

(c) What is the time complexity to find the best split using entropy assuming binary features? / continuous features?

O(nd) for binary classification

O($n^2 d$) for continuous regression

2(a)

$\mathrm{L(D)} = \frac{1}{|D|} \sum_{(\vec{x_j}, y_j) \in D} (y_j - \frac{1}{|S|} \sum_{(\vec{x_i}, y_i) \in D} y_j)^2$

$= \frac{1}{|D|} \sum_{(\vec{x_j}, y_j) \in D} [{y_j}^2 - 2y_j \ \frac{1}{|S|} \sum_{(\vec{x_i}, y_i) \in D} y_j + (\frac{1}{|S|} \sum_{(\vec{x_i}, y_i) \in D} y_j)^2]$

$\mathrm{L(S)} = \frac{1}{|S|} \sum_{(\vec{x_i}, y_i) \in D} (y_i - p)^2$

$L'(S) = \frac{1}{|S|}\sum_{(\overrightarrow{x_i},y_{i)\in D}}(y_i - p) \quad (-1) = 0$

$\mathrm{p} = \frac{\sum_{(\overrightarrow{x_i},y_{i)\in D}} y_i}{|S|}$ $(average)$ and $L''(S) > 0$

$\mathrm{L(S): convex, p} = \frac{\sum_{(\overrightarrow{x_i},y_{i)\in D}} y_i}{|S|}$, $L(S) achieves\ global\ minimum\ at\ the\ point$

2 (b)

(i)

L: $x_1, x_2, \dots, x_i$

R: $x_{i+1}, x_{i+2}, \dots, x_n$

$\vec{y}_L = \frac{1}{i}\sum_{j=1}^{i} y_i$

$\vec{y}_R = \frac{1}{n-i}\sum_{j=i+1}^{n} y_i$

(ii)

$\mathcal{L}_L^i: \frac{1}{i}\sum_{j=1}^{i}(y_i - \vec{y}_L)^2$

$\mathcal{L}_R^i: \frac{1}{n-i}\sum_{j=1+1}^{n}(y_i - \vec{y}_R)^2$

$\mathrm{O(n-i)}$

(iii)

$\vec{y}_L^{i+1} = \frac{1}{i+1}\sum_{j=1}^{i+1} y_j = \frac{1}{i+1}\sum_{j=1}^{i} y_j + y_{i+1} = \frac{1}{i+1}\left(i\vec{y}_L^i + y_{i+1}\right)$

(iv)

$\mathcal{L}_L^{i+1} = \frac{1}{i+1}\sum_{j=1}^{i+1}(y_i - \vec{y}_L^{i+1})^2 = \frac{1}{i+1}\sum_{j=1}^{i+1}\left(y_i^2 - 2y_i\vec{y}_L^{i+1} + \left(\vec{y}_L^{i+1}\right)^2\right)$

$s^i = \sum_{j=1}^{i} y_i^2$

$y_i\vec{y}_L^{i+1} = \frac{1}{i+1}\left(s^i + y_{i+1}^2\right) - \frac{2y_L^{i+1}}{i+1}\sum_{j=1}^{i+1} y_j$

$= s^i - 2\vec{y}_L^{i+1}\vec{y}_L^{i+1} + (\vec{y}_L^{i+1})^2$

$= s^i - (\vec{y}_L^{i+1})^2$

(v)

$s_{i+1} = \sum_{j=1}^{i+1} y_i^2 = s_i + y_{i+1}^2$

For each c_i: O(1) -> O(nd)

$O(n) + O(1)O(n) = O(n)$

$\mathrm{O}(\mathrm{d}\ \ \mathrm{n} \log n) + O(dn) = O(d\ \ n \log n)\ \ better\ \ than\ \ O(dn^2)$