02/14/2017
Speaker: Haomiao Meng

Multicategory classification

Data: $X_1, \cdots, X_n, Y_1, \cdots, Y_n \in \{1, 2, \cdots K\}$

$Y \sim$ Multinomial distribution.

Goal: Want a $\delta(X)$ to minimize

$$P(Y \neq \delta(X)) = \mathbb{E}[\mathbb{1}_{Y \neq \delta(X)}] = \mathbb{E}[\mathbb{E}(\mathbb{1}_{Y \neq \delta(X)} | X)]$$
$$= \mathbb{E}[p_1 \mathbb{E}(\mathbb{1}_{\delta(X) \neq 1} | X) + \cdots + p_K \mathbb{E}(\mathbb{1}_{\delta(X) \neq K} | X)]$$

where $p_j(X) = P(Y = j | X)$

$$\delta^*(X) = \underset{j}{\operatorname{argmax}}\ p_j(X)$$

$\delta(X)$ is a fcn mapping X to $\{1, 2, \cdots, K\}$

1-D Example $K = 2$.

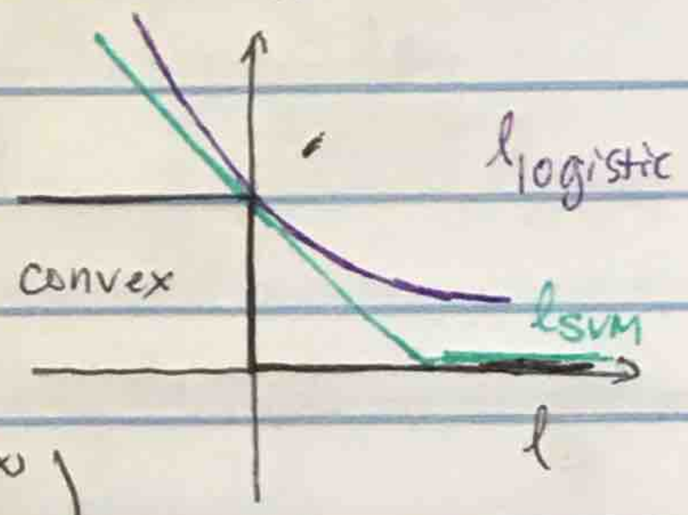
Data: $X_1, \cdots, X_n$, $Y \in \{1, -1\}$.

Goal: Want $\delta(X) = \text{sign}(f(x))$ to minimize

$$P(Y \neq \delta(X)) = \mathbb{E}(\mathbb{1}_{Y \neq \text{sign}(f(x))}) = \mathbb{E}(\mathbb{1}_{Yf(x) < 0}).$$

$\Rightarrow$ Want $\delta(X)$ to minimize. $\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{y_i f(x_i) < 0}$.

where $f(x) = X^T \beta$.



loss fcn examples:

0-1 loss: $l(Y, f(x)) = \mathbb{1}_{Yf(x) < 0}$ → not convex

SVM: $l_{SVM}(Y, f(x)) = (1 - Yf(x))_+$

logistic: $l_{logistic}(Y, f(x)) = \ln(1 + e^{-Yf(x)})$

for SVM loss:

$$f^*(x) = \underset{f}{\operatorname{argmin}}\ \mathbb{E}[l_{SVM}(Y, f(x))] = \underset{f}{\operatorname{argmin}}\ \mathbb{E}[(1 - Yf(x))_+]$$

claim: $\delta^*(x) = \text{sign}(f^*(x))$

quick proof: $\mathbb{E}[(1 - Yf(x))_+] = \mathbb{E}[\mathbb{E}(1 - Yf(x))_+ | X]$

$$= \mathbb{E}[p_1 (1 - f(x))_+ + p_2 (1 + f(x))_+]$$

$$\therefore f^*(x) = \begin{cases} 1 & \text{if } p_1 \geq p_2 \\ -1 & \text{if } p_1 < p_2 \end{cases} \Rightarrow \delta^*(x) = \underset{\delta}{\operatorname{argmin}}\ \mathbb{E}(\mathbb{1}_{Y \neq \delta(X)}) = \text{sign}(f^*(x))$$

multicategory case

$f(x) = x^T\beta$ is not invaient among permutations of class labels.

Method 1:

let $\delta(x) = \underset{j}{\operatorname{argmax}}\ f_j(x)$, $\underline{f} = \begin{bmatrix} f_1(x) \\ \vdots \\ f_K(x) \end{bmatrix}$ s.t. $\sum_j f_j = 0$.

put a constrain to avoid identifiability issue.

$P(Y \neq \delta(X)) = \mathbb{E}[\mathbb{1}_{Y \neq \underset{j}{\operatorname{argmax}} f_j(X)}]$

where the loss fcn $\mathbb{1}_{Y \underset{j}{\operatorname{argmax}} f_j(X)}$ is not convex

could replace the loss by

① Naive SVM: $l_S = (Y, \underline{f}(X)) = (1 - f_Y(X))_+$

the following shows that when $K=2$, the l_S is the same as the l_{SVM} defined above:

$l_{SVM} = \mathbb{E}(1 - Yf(X))_+ = \mathbb{E}[p_1(1-f(X))_+ + p_2(1+f(X))_+]$

$l_S = \mathbb{E}(1 - f_Y(X)) = \mathbb{E}[p_1(1-f_1(X))_+ + p_2(1-f_{-1}(X))_+]$

since $f_1 + f_{-1} = 0 \Rightarrow f_{-1} = -f_1 \Rightarrow l_{SVM} = l_S$.

then $f^*(x) = \underset{f}{\operatorname{argmax}}\ \mathbb{E}[l_S(Y, \underline{f}(X))] = \begin{cases} -(K-1), & j = \underset{j}{\operatorname{argmin}}\ p_j \\ 1, & \text{OW} \end{cases}$.

$\delta^*(x) = \underset{j}{\operatorname{argmax}}\ f_j^*(x) = \{1, 2, \dots, K\} \setminus \{\underset{j}{\operatorname{argmin}}\ p_j\}$.

not consistent.

② (Lee et al. (2004)): $l(Y, \underline{f}(X)) = \sum_{j \neq Y} (1 + f_j(x))_+$.

$f_j^*(x) = \begin{cases} K-1, & j = \underset{j}{\operatorname{argmax}}\ p_j \\ -1, & \text{OW} \end{cases}$.

Claim: $\delta^*(x) = \underset{j}{\operatorname{argmax}}\ f_j^*(x) = \underset{j}{\operatorname{argmax}}\ p_j$

quick proof: (next page)

Quick proof of the above claim:

① $\underline{f}^*$ satisfies $f_i^* \geq -1 \Rightarrow l(Y, \underline{f}(X)) = \sum_{j \neq Y} (1 + f_j(X))$

② multicategory SVM:

$$\min \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq Y_i} (1 + f_j(X_i))_+ \quad \text{w/ } \sum f_j = 0, f_j \geq -1$$

$$\Leftrightarrow \min -\sum_i f_{Y_i}(X_i) + \lambda \|\underline{f}\|_2^2 \quad \text{s.t. } \sum f_j = 0, f_j \geq -1.$$

$$\sum_{j \neq Y_i} (1 + f_j(X_i))_+ = \sum_{j \neq Y_i} (1 + f(X_i))$$

$$= K - 1 + \sum_{j \neq Y_i} f_j(X_i)$$

$$= K - 1 - f_{Y_i}(X_i)$$

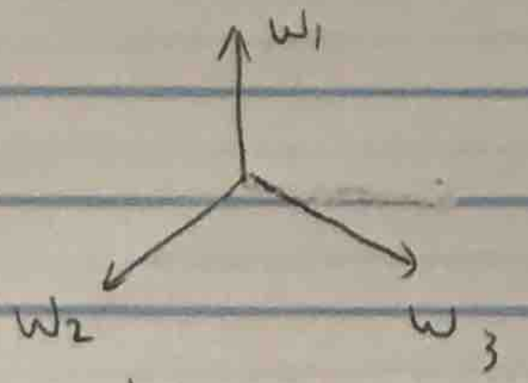
Method 2

Data: $\underline{X}_1, \cdots, \underline{X}_n$, $Y_1, \cdots, Y_n$, $W_{Y_1}, \cdots, W_{Y_n}$.

where W_{Y_i}'s form a simplex.

Examples of W_{Y_i}'s

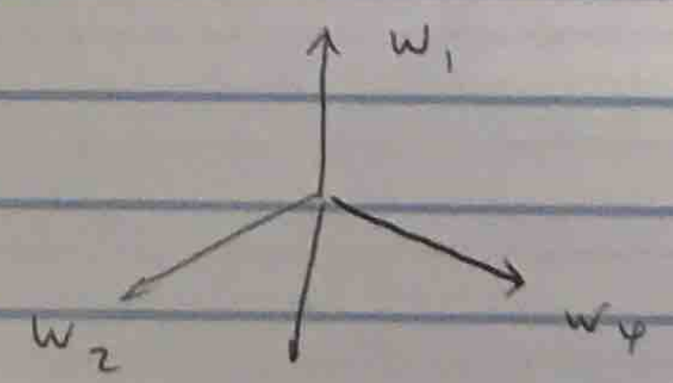
$K = 3$, 2-D



angels b/w each two are same.

the. $w_1 + w_2 + w_3 = 0$.

$K = 4$, 3-D



$w_1 + w_2 + w_3 + w_4 = 0$

then could let

$$\delta(X) = \arg\min \angle(w_j, \underline{f}(X)), \text{ where } \underline{f}(X) = \begin{bmatrix} f_1 \\ \vdots \\ f_{K-1} \end{bmatrix}$$

Since the angel is not easy to interpret, could change to inner product or norm:

e.g. $\delta(X) = \arg\min \langle w_j, \underline{f}(X) \rangle$ or $\delta(X) = \arg\min \|w_j - \underline{f}(X)\|^2$

let $\delta(x) = \text{argmax} \langle W_i, \underline{f}(x) \rangle$

$P(Y \neq \delta(x)) = \mathbb{E}(\mathbb{1}_{Y \neq \text{argmax} \langle W_j, \underline{f}(x) \rangle})$

$\min_f \mathbb{E}[l(Y, \underline{f}(x))] \Rightarrow^{j} \min_f \mathbb{E}[l(\langle W_Y, f(x) \rangle)]$

thm If l is strictly decreasing and convex, then $f^*(x)$ is fisher consistent.

could minimize $\frac{1}{n} \sum_{i=1}^{n} l(\langle W_{Y_i}, \underline{f}(x_i) \rangle) + \lambda \|\underline{f}\|_2^2$

to control the "length" of f.

there is an underlying constrain:

Since $W_1 + W_2 + W_3 = 0 \Rightarrow \langle W_1, \underline{f} \rangle + \langle W_2, \underline{f} \rangle + \langle W_3, \underline{f} \rangle = 0$

General Steps for Multicategory classification:

Step 1: choose form of δ

2: choose loss fcn

3: choose estimation procedure.