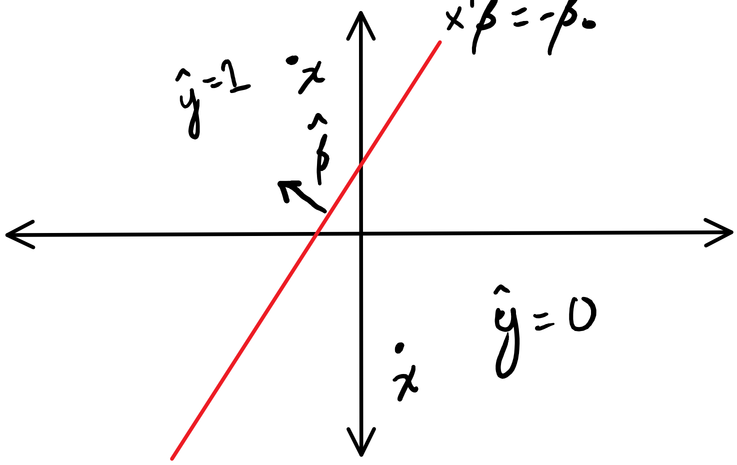
Margin Based Methods

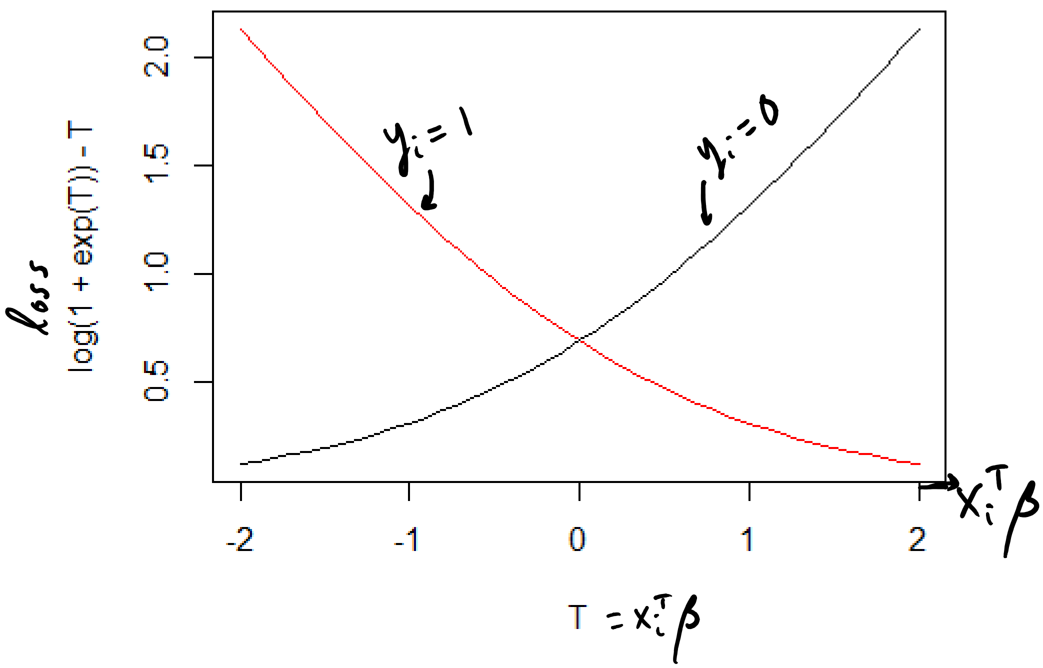
Tuesday, April 25, 2017 4:22 PM

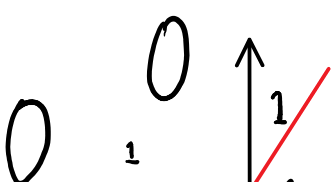
Predict for logistic regression and LDA : $\hat{y} = 1\{x^T\hat{\beta} + \hat{\beta}_0 \geq 0\}$



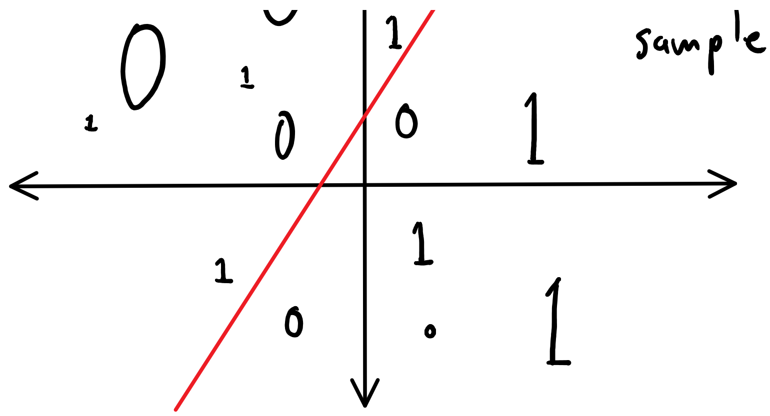
Empirical Risk Minimization

$$\min_{\beta} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, x_i, \beta) \rightarrow -y_i x_i^T\beta + \log(1+e^{x_i^T\beta})$$





Loss weighted sample



Recall 0-1 loss $l(y_i, x_i, \beta) = 1\{y_i = \hat{y}_i\}$ and re-encode

$y_i \leftarrow 2y_i - 1$ so that $y_i \in \{-1, 1\}$, and $\hat{y}_i = \text{sign}(x_i^T\beta)$

Error if $y_i \cdot \hat{y}_i = 1 \Leftrightarrow y_i \cdot x_i^T\beta > 0$ so

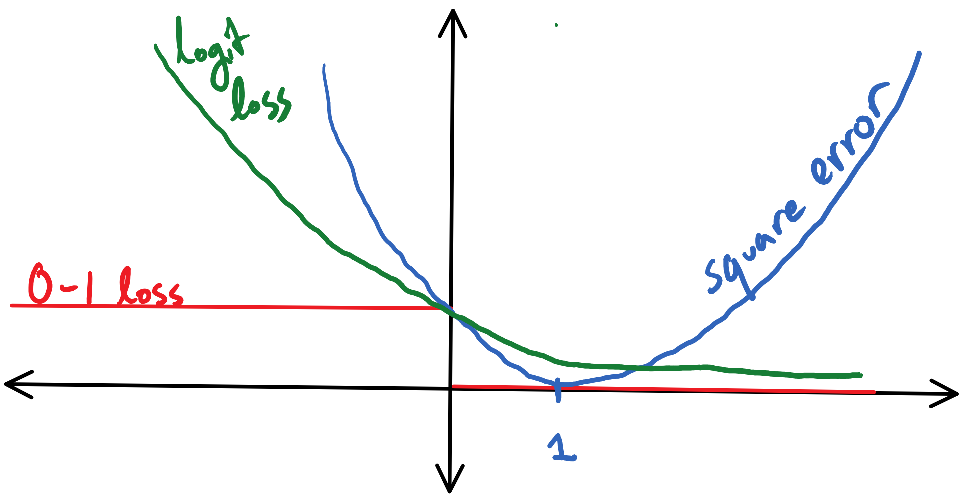
$$l_{0/1}(y_i, x_i, \beta) = 1\{y_i \cdot x_i^T\beta \leq 0\}$$

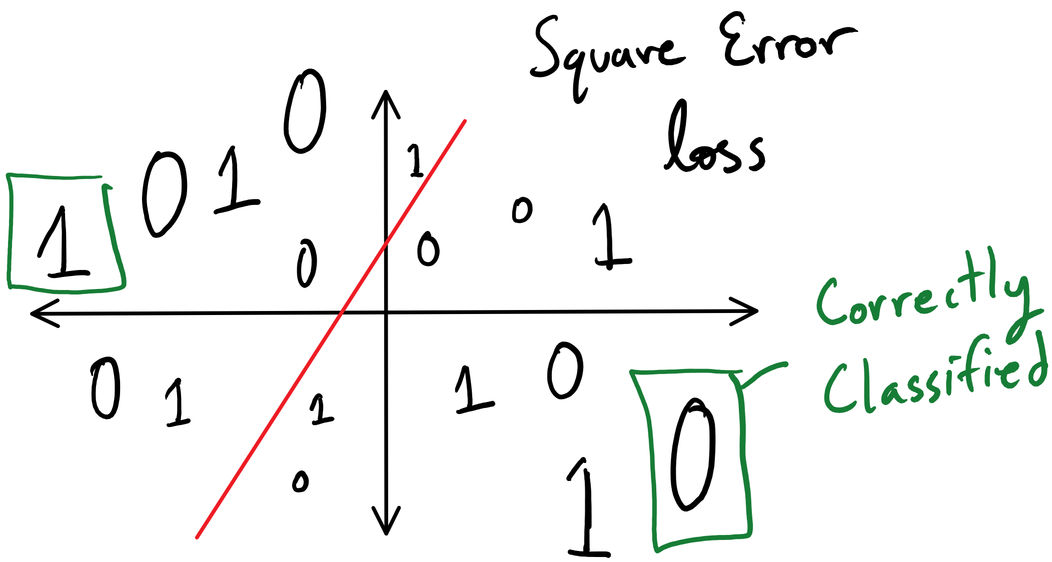
and
$$l_{logit}(y_i, x_i, \beta) = \begin{cases} \log(1 + e^{x_i^T\beta}), & y_i = -1 \\ \log(1 + e^{-x_i^T\beta}), & y_i = 1 \end{cases}$$

$$= \log(1 + e^{-y_i x_i^T\beta})$$

You can also use square error loss,

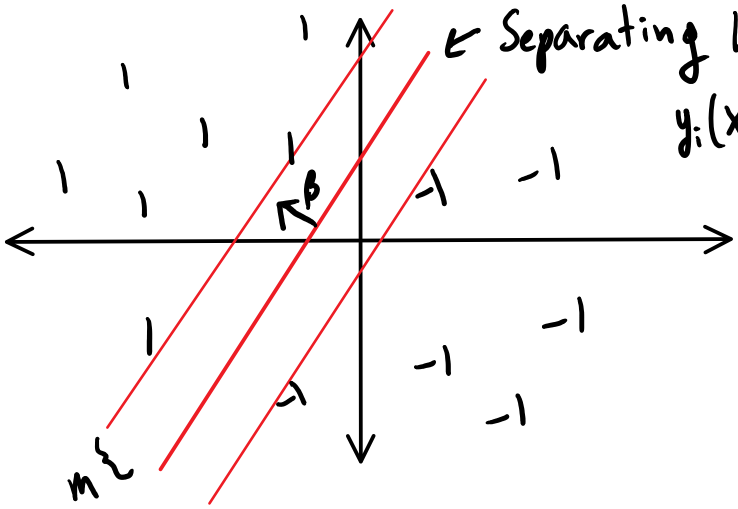
$$l_2(y_i, x_i, \beta) = (y_i - x_i^T\beta)^2 = (1 - y_i x_i^T\beta)^2$$


logit loss
square error
0-1 loss
1


Square Error loss
Correctly Classified

Support Vector Machines

Wednesday, April 26, 2017 12:04 PM



← Separating hyperplane
$y_i(x_i^T\beta+\beta_0)\geq 0$

Max-margin separating hyperplane

$$\max_{\beta,\beta_0:\|\beta\|=1} M \quad \text{s.t. } y_i(x_i^T\beta+\beta_0)\geq M \ \forall i$$

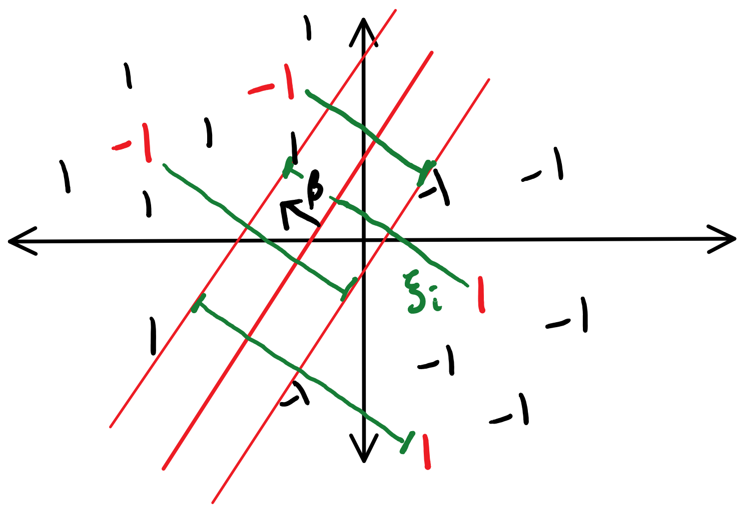
$$\equiv \max_{\beta,\beta_0} M \text{ s.t. } \frac{1}{\|\beta\|} y_i(x_i^T\beta+\beta_0)\geq M \ \forall i$$

↳ can scale β arbitrarily so set $\|\beta\| = \frac{1}{M}$

$$\equiv \min_{\beta,\beta_0} \frac{1}{2}\|\beta\|_2^2 \text{ s.t. } y_i(x_i^T\beta+\beta_0)\geq 1 \ \forall i$$

↳ solution only dependent on "support vectors"

!!! Feasible only if linearly separable !!!



If not linearly separable then add "slack variable" ξ_i

$$\min \|\beta\| \text{ s.t. } \quad y_i(x_i^T\beta+\beta_0) \geq 1-\xi_i \quad \forall i$$
$$\xi_i \geq 0 \quad \sum_i \xi_i \leq C$$

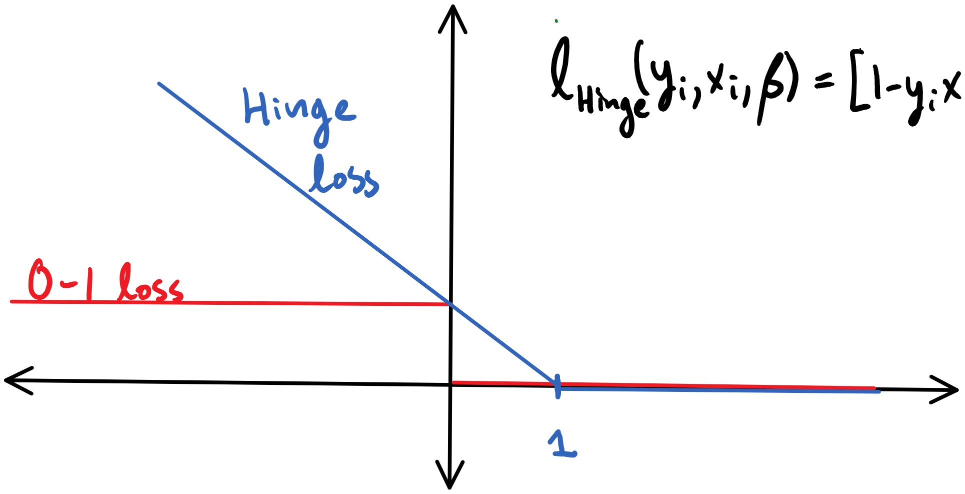
Lagrangian

$$\min_{\beta_0,\beta,\xi_i\geq 0} \|\beta\|_2^2 + \gamma \sum_i \xi_i \quad \text{s.t. } y_i(x_i^T\beta+\beta_0) \geq 1-\xi_i \ \forall i$$

$$\equiv \min_{\beta,\beta_0} \sum_i \left[1-y_i x_i^T\beta\right]_+ + \lambda \|\beta\|_2^2$$

where $a_+ = \begin{cases} a, & a \geq 0 \\ 0, & a < 0 \end{cases}$.

$$\ell_{Hinge}(y_i, x_i, \beta) = [1 - y_i x_i^T \beta]_+$$



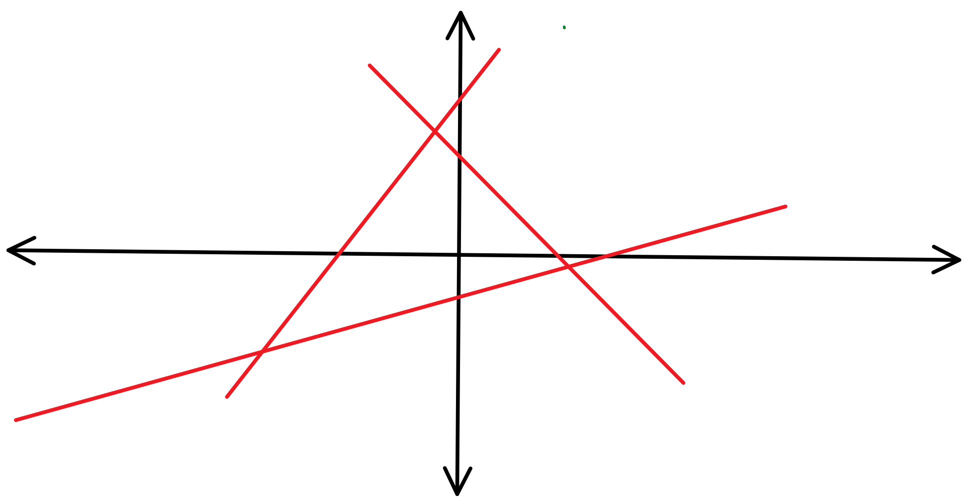
Multiclass Classification

Wednesday, April 26, 2017 5:15 PM

Encode K classes: $y_i \in \{0,1\}^K$

$$\ell_{0,1}(y_i, \hat{y}_i) = 1 - y_i^T \hat{y}_i$$

Multiple linear separators: $\{\beta_k\}_{k=1}^K$



$$\hat{y}_j = \begin{cases} 1, & j = \arg\max_k \beta_k^T x \\ 0, & \text{otherwise} \end{cases}$$

Soft-argmax $\quad z \in \mathbb{R}^K$ is vector of scores

$$s(z) = \left(\frac{e^{z_1}}{\sum_k e^{z_k}}, \frac{e^{z_2}}{\sum_k e^{z_k}}, \dots, \frac{e^{z_K}}{\sum_k e^{z_k}} \right)$$

Replace $\hat{y}_i = s(x_i^T \hat{\beta}_1, \dots, x_i^T \hat{\beta}_K)$ then

$y_i^T \hat{y}_i = e^{x_i^T \hat{\beta}_j}$ for $y_{ij} = 1$

$$y_i^T \hat{y}_i = \frac{e^{x_i^T \hat{\beta}_j}}{\sum_k e^{x_i^T \hat{\beta}_k}} \quad \text{for} \quad y_{ij} = 1 .$$

$$= \frac{e^{x_i^T \overbrace{(\hat{\beta}_j - \hat{\beta}_K)}^{\text{rename } \hat{\beta}_j}}}{1 + \sum_{k=1}^{K-1} e^{x_i^T(\hat{\beta}_k - \hat{\beta}_K)}} = \mathbb{P}\{\text{Class} = j \mid x\}$$

for logistic model

$$\log \frac{\mathbb{P}\{\text{Class } j \mid x\}}{\mathbb{P}\{\text{Class } K \mid x\}} = x^T \hat{\beta}_j .$$

- multiclass SVM predicts class with max margin / smallest slack var.
- Confusion matrix is $K \times K$.