lecture 4. SVM & Regularization, Logistic regression & Maximum likelihood

PRIMAL-SVM with Slack

$$\min_{(\underline{\theta}, \theta_0, \underline{\xi}) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^n} \frac{1}{2}\|\underline{\theta}\|^2 + C\sum_{t=1}^{n} \xi_t$$

such that $\xi_t \geq 0$ $\quad y_t(\langle x_t, \underline{\theta}\rangle + \theta_0) \geq 1 - \xi_t$ $\quad t = 1, 2, \ldots, n$

$C > 0$: constant $\quad [C \uparrow +\infty \Rightarrow \xi_t \downarrow 0]$

ξ_t : slack parameters (nonnegative)

quadratic program:

$\rightarrow \min \frac{1}{2} x^T H x + c^T x$

s.t. $Ax \leq b$.

Machine Learning Problem:

$$\min_{\underline{\theta}, \theta_0} \frac{1}{n}\sum_{t=1}^{n} \text{Loss}(y_t, f_{\underline{\theta}, \theta_0}(x_t)) + \lambda R(\underline{\theta})$$

- parameters
- target/label
- prediction
- Empirical loss
- Regularizer
- Regularization term

Transformation of PRIMAL-SVM with Slack.

(rewritten)

$$\min_{(\underline{\theta}, \theta_0, \underline{\xi}) \in \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^n} \frac{1}{2}\|\underline{\theta}\|^2 + C\sum_{t=1}^{n} \underbrace{(1 - y_t(\langle \underline{\theta}, x_t\rangle + \theta_0))^+}_{\xi_t}$$

viewed as 'agreement'

$(a)^+ = \max\{a, 0\}$

$\hat{\xi}_t = [1 - y_t(\langle x_t, \underline{\theta}\rangle + \theta_0)]^+$

(i) If no slack, $y_t(\langle x_t, \underline{\theta}\rangle + \theta_0) \geq 1 \Rightarrow \hat{\xi}_t = 0$

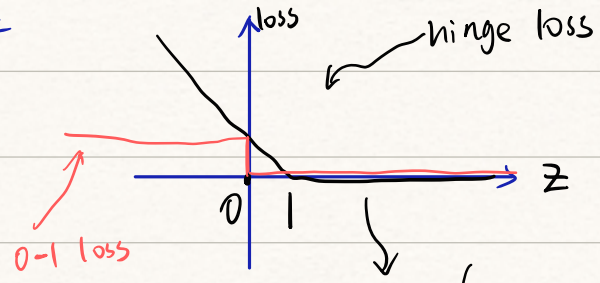
(ii) If there is a slack, $y_t(\langle x_t, \underline{\theta}\rangle + \theta_0) < 1 \Rightarrow \hat{\xi}_t > 0$

$\xi_t = 1 - y_t(\langle x_t, \theta \rangle + \theta_0)$

$\text{Loss}_h(z) = (1-z)^+$

↓

hinge loss



if $y_t(\langle x_t, \theta \rangle + \theta_0) \geq 1 \rightsquigarrow$ perfectly classified

↑ no loss!

$C \sum_{t=1}^{n} (1 - y_t(\langle x_t, \theta \rangle + \theta_0))^+$

$= C \sum_{t=1}^{n} \text{Loss}_h(y_t(\langle x_t, \theta \rangle + \theta_0))$

if $y_t(\langle x_t, \theta \rangle + \theta_0) < 1$

↳ { classified correctly but not "well".
misclassified

⇓

linear loss!

$\Rightarrow$ PRIMAL-SVM with slack

$\Leftrightarrow \min_{(\theta, \theta_0) \in \mathbb{R}^d \times \mathbb{R}} C \sum_{t=1}^{n} \text{Loss}_h(y_t(\langle x_t, \theta \rangle + \theta_0)) + \frac{1}{2}\|\theta\|^2$ ↘ equivalent (why?)

$\Leftrightarrow \min_{(\theta, \theta_0) \in \mathbb{R}^d \times \mathbb{R}} \sum_{t=1}^{n} \text{Loss}_h(y_t(\langle x_t, \theta \rangle + \theta_0)) + \lambda \|\theta\|^2$

Empirical Loss

↓

$\lambda = \frac{1}{2C}$

Regularization term.

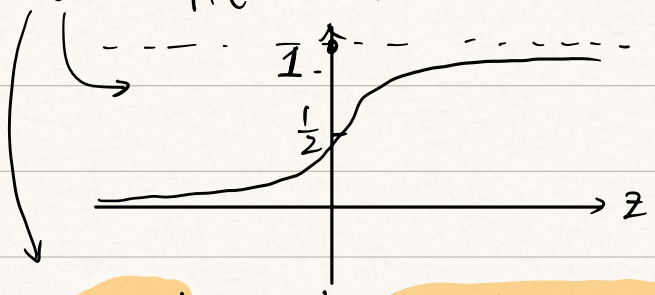
Logistic Regression

↳ assign a probability distribution over the two labels $\{+1, -1\}$ s.t labels far from the DB (and correctly labelled) are "more likely" to be correct!.

↳ model our labels by "confidence"

$P(y=1 \mid x, \theta, \theta_0) = g(\theta^T x + \theta_0)$

$g(\cdot): \mathbb{R} \rightarrow (0,1)$ $\quad g(z) = \frac{1}{1+e^{-z}}$ logistic function



have special property: $1 - g(z) = g(-z)$

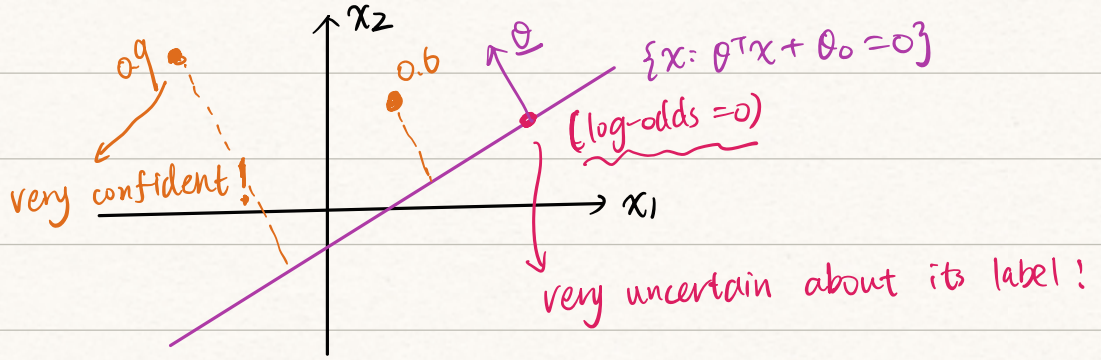
Question: where does $g(z)$ comes from?

→ Consider the log-odds: (linear function of inputs $\underline{x}$)

$$\log \frac{P(y=1 \mid \underline{x}, \underline{\theta}, \theta_0)}{P(y=-1 \mid \underline{x}, \underline{\theta}, \theta_0)} = \underline{\theta}^T \underline{x} + \theta_0$$

Note: if $P(y=1 \mid \cdots) = P(y=-1 \mid \cdots) = \frac{1}{2}$.

then: $\underline{\theta}^T \underline{x} + \theta_0 = 0 \Rightarrow$ we have decision boundary



$$P(y=-1 \mid x, \theta, \theta_0) = 1 - P(y=1 \mid x, \theta, \theta_0)$$

$$= 1 - g(\theta^T x + \theta_0)$$

$$= g(-(\theta^T x + \theta_0))$$

from this observation.

$$\Rightarrow P(y \mid x, \theta, \theta_0) = g(y(\theta^T x + \theta_0))$$

$$= \begin{cases} g(\theta^T x + \theta_0), & y=1 \\ g(-(\theta^T x + \theta_0)), & y=-1 \end{cases}$$

logistic regression → a probabilistic way for us to model!

Question: How do we train LR model?

⇔ How to learn $(\underline{\theta}, \theta_0)$ given $\mathcal{D} = \{(x_t, y_t) : 1 \le t \le n\}$.

↳ Suprisingly, its update process is very similar to Perceptron!

↳ Method: MLE to learn $(\underline{\theta}, \theta_0)$.

Likelihood: $L(\underline{\theta}, \theta_0 | \mathcal{D}) = \prod_{t=1}^{n} P(y_t | x_t, \underline{\theta}, \theta_0)$

$(\hat{\underline{\theta}}, \hat{\theta}_0) = \underset{(\underline{\theta}, \theta_0)}{\text{argmax}}\ L(\underline{\theta}, \theta_0 | \mathcal{D})$. → Many beautiful properties! (section 4).

log-likelihood

$$\Leftrightarrow (\hat{\underline{\theta}}, \hat{\theta}_0) = \underset{(\underline{\theta}, \theta_0)}{\text{argmax}} \sum_{t=1}^{n} \log P(y_t | \underline{x}_t, \underline{\theta}, \theta_0)$$

$$= \underset{(\underline{\theta}, \theta_0)}{\text{argmin}} \sum_{t=1}^{n} -\log P(y_t | x_t, \underline{\theta}, \theta_0)$$

$$= \underset{(\underline{\theta}, \theta_0)}{\text{argmin}} \sum_{t=1}^{n} -\log\left(g\left(y_t\left(\underline{\theta}^T \underline{x}_t + \theta_0\right)\right)\right)$$

Note: $g(z) = \frac{1}{1+\exp(-z)}$

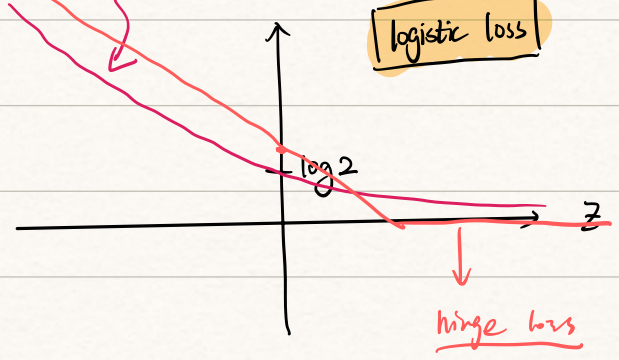
$$= \underset{(\underline{\theta}, \theta_0)}{\text{argmin}} \sum_{t=1}^{n} \log\left(1 + \exp\left(-y_t(\underline{\theta}^T \underline{x}_t + \theta_0)\right)\right)$$

↑ agreement z

$$= \underset{(\underline{\theta}, \theta_0)}{\text{argmin}} \sum_{t=1}^{n} \text{loss}_{\log}\left(y_t(\underline{\theta}^T \underline{x}_t + \theta_0)\right)$$

↓ No regularization term

★ $\text{Loss}_{\log}(z) = \log(1 + \exp(-z))$



Rmk: The logistic loss (as with the hinge loss) depends only on the agreement $y_t(\langle \underline{\theta}, \underline{x}_t \rangle + \theta_0)$ → it can be negative

反映了 { label y_t
prediction (linear) $\langle \underline{\theta}, x_t \rangle + \theta_0$ 的相似程度.

Analysis:

Start from the likelihood.

↓

log-likelihood

↓

Logistic Loss function $\left(\text{Loss}_{\log}(z) = \log(1+\exp(-z)) \right)$

$$\underset{(\underline{\theta}, \theta_0)}{\text{argmin}} \sum_{t=1}^{n} \text{loss}_{\log}\left(y_t(\underline{\theta}^T x_t + \theta_0) \right)$$

$y_t(\underline{\theta}^T x_t + \theta_0)$ → agreement.

↑

Optimization Problem

We want to optimize over $(\underline{\theta}, \theta_0)$.

↓

$$f(\underline{\theta}, \theta_0) := \sum_{t=1}^{n} \log\left(1+\exp\left(-y_t(\underline{\theta}^T x_t + \theta_0)\right)\right) = \sum_{t=1}^{n} f_t(\underline{\theta}, \theta_0)$$

↳ think about the gradient with $\underline{\theta}$ & θ_0

① $\dfrac{d}{d\theta_0} f_t(\underline{\theta}, \theta_0) = \dfrac{d}{d\theta_0} \log\left[1+\exp(-y_t(\theta^T x_t + \theta_0))\right]$.

$$= \frac{-y_t \exp(-y_t(\theta^T x_t + \theta_0))}{1+\exp(-y_t(\theta^T x_t + \theta_0))}$$

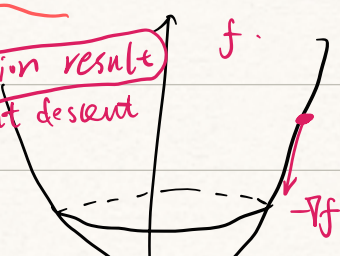
$$= -y_t\left[1 - P(y_t \mid x_t, \underline{\theta}, \theta_0)\right]$$

$P(y_t \mid x_t, \underline{\theta}, \theta_0) = g(y_t(\theta^T x_t + \theta_0)) = \dfrac{1}{1+\exp(-y_t(\theta^T x_t + \theta_0))}$

② Similarly, $\dfrac{d}{d\underline{\theta}} f_t(\underline{\theta}, \theta_0) = -y_t x_t \left[1 - P(y_t \mid x_t, \underline{\theta}, \theta_0)\right]$. Beautiful!

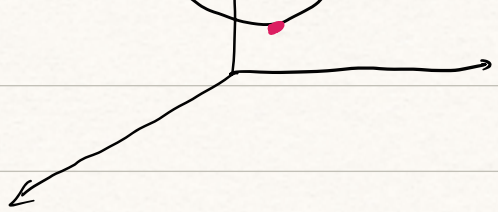
soft classification result

gradient descent



Gradient Descent: (GD)

$$x^{(k+1)} = x^{(k)} - \eta_k \nabla f(x^{(k)})$$



To minimize sum of logistic losses.

move in the negative gradient direction

↳ (Stochastic GD) → SGD (update term by term).

Updates:

① $$\underline{\theta}^{(k+1)} = \underline{\theta}^{(k)} - \eta_k \nabla f_t(\underline{\theta}^{(k)}, \theta_0^{(k)})$$ → stochastic choose the t-th data sample!

$$= \underline{\theta}^{(k)} + \eta_k y_t \underline{x}_t \left[1 - P(y_t | \underline{x}_t, \theta_0^{(k)}, \underline{\theta}^{(k)}) \right].$$

② $$\theta_0^{(k+1)} \leftarrow \theta_0^{(k)} + \eta_k y_t \left[1 - P(y_t | \underline{x}_t, \underline{\theta}^{(k)}, \theta_0^{(k)}) \right].$$

the probability of making mistake for current $(\underline{\theta}^{(k)}, \theta_0^{(k)})$

Focus on the update for $\underline{\theta}$ (set $\eta = 1$)

$$\underline{\theta} \leftarrow \underline{\theta} + y_t \underline{x}_t \left[1 - P(y_t | \underline{x}_t, \underline{\theta}, \theta_0) \right].$$

→ the probability of correct prediction

probability of making a mistake when we predict the label $\underline{x}_t$ using current parameters $(\underline{\theta}, \theta_0)$ (How Wrong!)

Perceptron Alg. → update while making a mistake

→ $\underline{\theta} \leftarrow \underline{\theta} + y_t \underline{x}_t$ if prediction of y_t from data sample $\underline{x}_t$ is wrong!

Note: These two updates are analogue

And the LR updates are "soft" !!!

↳ using the probability.

Rmk: In LR Updates, updates are made in proportion to the prob. of making mistakes!

☆ if a data sample is more likely

to be incorrect, then the update of $\underline{\theta}$ (θ_0) is bigger with respect to $y_t x_t$ (y_t)

Necessary condition for optimality of $(\underline{\theta}^*, \theta_0^*)$

$$\begin{cases} \frac{d}{d\theta_0} \ell(\theta, \theta_0 | \mathcal{D}) = \sum_{t=1}^{n} -y_t [1 - P(y_t | x_t, \underline{\theta}, \theta_0)] \Big|_{\theta_0 = \theta_0^*} = 0 & (+) \\ \frac{d}{d\underline{\theta}} \ell(\theta, \theta_0 | \mathcal{D}) = \sum_{t=1}^{n} -y_t x_t [1 - P(y_t | x_t, \underline{\theta}, \theta_0)] \Big|_{\underline{\theta} = \underline{\theta}^*} = \underline{0} & (++) \end{cases}$$

$\ell(\theta, \theta_0 | \mathcal{D})$ → log-likelihood

Here, $y_t \in \{-1, 1\}$

↓ map: $\tilde{y}_t = \frac{1+y_t}{2}$ $\begin{cases} -1 \to 0 \\ +1 \to +1 \end{cases}$

$\tilde{y}_t \in \{0, 1\}$

Claim: (+) reduces to

$$\sum_{t=1}^{n} [\tilde{y}_t - P(1 | \underline{x}_t, \underline{\theta}^*, \theta_0^*)] = 0$$

$e_t \in \mathbb{R}$

→ Can be proved as follows

if $y_t = 1$, $\tilde{y}_t = 1$ ✓

if $y_t = -1$, $\tilde{y}_t = 0$

⇓

$y_t - y_t P(y_t | x_t, \underline{\theta}^*, \theta_0^*)$

$= -1 + P(-1 | x_t, \underline{\theta}^*, \theta_0^*)$

$= P(1 | x_t, \underline{\theta}^*, \theta_0^*)$

(++) reduces to

$$\sum_{t=1}^{n} \underline{x}_t [\tilde{y}_t - P(1 | \underline{x}_t, \underline{\theta}^*, \theta_0^*)] = \underline{0}$$

↓

prediction error $e_t \in \mathbb{R}$

Can be { positive → '+' class, negative → '−' class

its abs. is the prediction goes wrong prob.

denote prediction error $\underline{e} = \begin{bmatrix} \tilde{y}_1 - P(1 | x_1, \underline{\theta}^*, \theta_0^*) \\ \vdots \\ \tilde{y}_n - P(1 | x_n, \underline{\theta}^*, \theta_0^*) \end{bmatrix}$

Claim: Any linear (affine) function of inputs $\underline{x}_t \in \mathbb{R}^d$ is orthogonal to prediction error $\underline{e}$.

$(\hat{\underline{\theta}}^T \underline{x}_t + \hat{\theta}_0)_{1\times n}$

Pf: $\hat{\theta}_0 \sum_{t=1}^{n} e_t + \langle \hat{\underline{\theta}}, \sum_{t=1}^{n} \underline{x}_t e_t \rangle = 0$ $\left(\hat{\theta}_0 \in \mathbb{R},\ \hat{\underline{\theta}} \in \mathbb{R}^d.\right)$

$$\Rightarrow \sum_{t=1}^{n} (\tilde{\theta_0} + \langle \hat{\theta}, \underline{x_t} \rangle)\, e_t = 0$$

any linear function of data $\rightarrow$ $\underline{d}$

prediction error $\rightarrow$ $\underline{e}$

$$\underline{d} = \begin{bmatrix} \hat{\theta_0} + \langle \hat{\theta}, \underline{x_1} \rangle \\ \vdots \\ \hat{\theta_0} + \langle \hat{\theta}, \underline{x_n} \rangle \end{bmatrix}$$

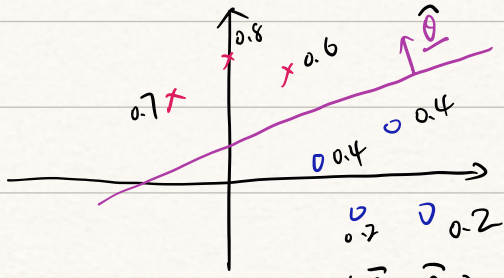
$\Rightarrow \langle d, e \rangle = 0$ $\rightarrow$ this actually means we cannot use data sample to improve our solution (minimize)

If training data $(x_t, y_t)_{t=1}^{n}$ is affinely separable.

$\Rightarrow \exists\, (\underline{\theta}^*, \theta_0^*)$ s.t $y_t(\langle \underline{x_t}, \underline{\theta}^* \rangle + \theta_0^*) \geq 0$ for all t.

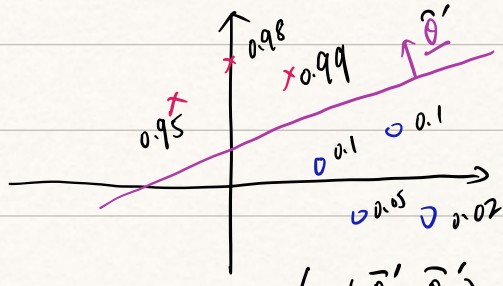
$\downarrow$ logistic regression

Question: is it good? $\rightarrow$ NO! (over-fitting)



$P(y=1 \mid \underline{x}, \hat{\underline{\theta}}, \hat{\theta_0})$

$(\hat{\underline{\theta}}, \hat{\theta_0})$ $\leftarrow$ learn from logistic model



$P(y=1 \mid \underline{x}, \hat{\underline{\theta}}', \hat{\theta_0}')$

$(\hat{\underline{\theta}}', \hat{\theta_0}') = (c \cdot \hat{\underline{\theta}}, c \cdot \hat{\theta_0})$ $c \gg 1$

this is because $g(z) = \frac{1}{1+e^{-z}}$ $\nearrow$ as $z \uparrow$

$$z = y_t(\langle x_t, \theta \rangle + \theta_0)$$

Rmk: Because dataset is linearly separable, there $\exists (\theta^*, \theta_0^*)$,

s.t $y_t(\langle x_t, \theta^* \rangle + \theta_0^*) \geq 0 \quad \forall t$

& $g(y_t(\langle x_t, \theta^* \rangle + \theta_0^*)) = \dfrac{1}{1+\exp(-\underbrace{y_t(\langle x_t, \theta^* \rangle + \theta_0^*)}_{\geq 0})}$

Therefore, we choose

$(c\theta^*, c\theta_0^*) \longrightarrow (\theta^*, \theta_0^*)$

the $g(z)$ will become larger (any)! (choose c big enough)

In logistic regression, if $\mathcal{L}$ is affinely separable, then $(\theta^*, \theta_0^*) \to \infty$!

↓

Remedy: Add Regularization:

↓

$$\underbrace{\frac{\lambda}{2}\|\theta\|^2}_{\text{Regularization term}} + \underbrace{\sum_{t=1}^{n} \text{Loss}_{\log}(y_t \langle x_t, \theta \rangle + \theta_0)}_{\text{Empirical logistic loss}}$$

Multiclass case: $g(z)$ generalized

↓

softmax function

① PRIMAL SVM with slackness and offset

$$\min_{\theta} \ \frac{1}{2}\|\theta\|^2 + C\sum \xi_t$$
$$s.t \quad y_t(\langle x_t, \theta\rangle + \theta_0) \geq 1-\xi_t,$$
$$\xi_t \geq 0$$

Equivalent ⇔

Regularization form: $z_t = y_t(\langle x_t, \theta\rangle + \theta_0)$ ↪ agreement

$$\min_{\theta} \sum_t \text{loss}_{\text{hinge}}(z_t) + \frac{1}{2C}\|\theta\|^2$$
$$\Leftrightarrow \min_{\theta} \sum_t (1-y_t(\langle x_t, \theta\rangle + \theta_0))^+ + \frac{1}{2C}\|\theta\|^2$$

Check: any (θ, θ_0, ξ) in (P)

Fix (θ, θ_0), notice that $\min \frac{1}{2}\|\theta\|^2 + C\sum \xi_t$

we should select the minimal ξ with (θ, θ_0).

$$\xi_t \rightsquigarrow \begin{cases} 1-y_t(\langle x_t, \theta\rangle + \theta_0) \leq \xi_t \\ \xi_t \geq 0 \end{cases}$$

$$\xi_t = \max\{0, 1-\text{agreement}\} = (1-\text{agreement})^+$$

$$\Rightarrow (P) \Leftrightarrow \min \frac{1}{2}\|\theta\|^2 + C\sum \xi_t$$
$$s.t \ \xi_t = (1-\text{agreement})^+$$

⇒ The Equivalence Has proved