Kernel Logistic Regression

JrPhy

Introduction

- Now we record the error data in ξ, then minimize $\boldsymbol{w}$ with the constrain,

$$\boxed{\min_{w}\left(\frac{1}{2}\|w\|^{2}+C\sum_{i}\xi_{i}\right)=\min_{w}\left(\frac{1}{2}ww^{T}+C\sum_{i}err_{i}\right)}$$

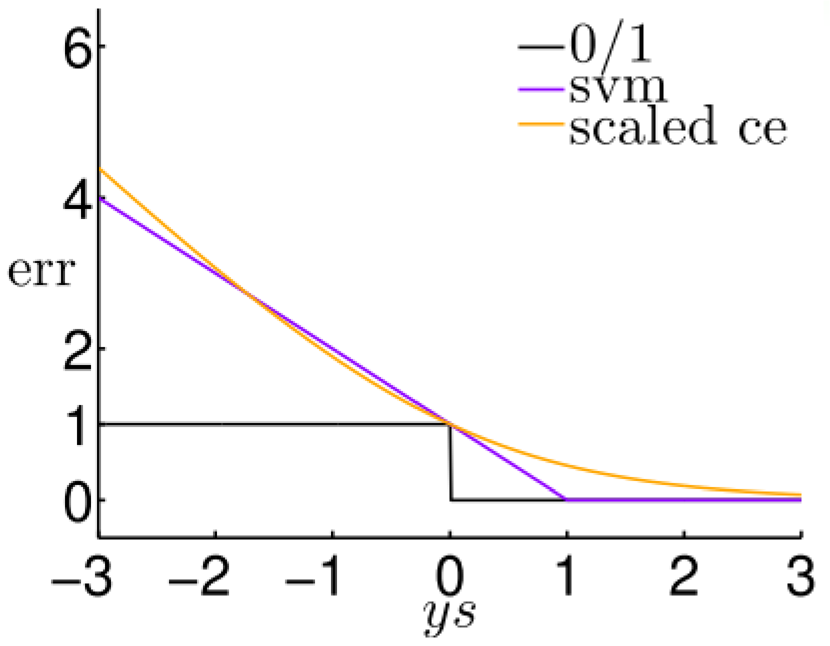
- This form is like the regularization

$$\boxed{\min_{w}\left(\frac{\lambda}{N}ww^{T}+\frac{1}{N}\sum_{i}err_{i}\right)}$$

- Now let’s connect the soft-margin SVM and logistic regression

Logistic regression → soft-margin

- The constrain of soft-margin SVM is
- $y(\boldsymbol{w}^{T}\boldsymbol{x} + \boldsymbol{b}) \geq 1 - \xi_i,\ \xi_i \geq 0 \rightarrow \xi_i \geq 1 - y(\boldsymbol{w}^{T}\boldsymbol{x}+b)$,
- this equivalent as $\max(1 - y(\boldsymbol{w}^{T}\boldsymbol{x}+b), 0)$. So the score of soft-margin SVM is
- $err_{SVM} = \max(1 - y(\boldsymbol{w}^{T}\boldsymbol{x}+b), 0)$
- The score of Logistic regression is
- $err_{Logistic} = \log 2(1 - y(\boldsymbol{w}^{T}\boldsymbol{x}+b))$, Let $s = \boldsymbol{w}^{T}\boldsymbol{x}+b$



Hsuan Tien Lin, mltech/205_handout

$$\boxed{\begin{array}{l} ys \rightarrow \infty,\ err_{SVM}\ \&\ err_{Logistic} \rightarrow 0 \\ ys \rightarrow -\infty,\ err_{SVM}\ \&\ err_{Logistic} \rightarrow \infty \end{array}}$$

Combine logistic regression and SVM

- So we can run SVM to get the b_{SVM} and $\mathbf{w}_{SVM}$, then input then in the logistic regression, but we can't get a target function with a probability distribution.
- Or we can set the b and $\mathbf{w}$ as initial condition of gradient descent to get the optimal b_{opt} and $\mathbf{w}_{opt}$, but it can not use kernel trick because of the nonlinear transform. So that we have to modify the score.
- The idea is running SVM with a kernel first, then take as a score $z = \mathbf{w}_{SVM}{}^{T}\mathbf{\Phi(x)}+b_{SVM}$ .then times A and plus B so that the form $Az+B$, so that the form is similar to logistic regression.

Combine logistic regression and SVM

- $g(x) = \theta(Az+B) = \theta(A(\boldsymbol{w}_{SVM}{}^{\boldsymbol{T}}\boldsymbol{\Phi}(\boldsymbol{x})+b_{SVM})+B)$

$$\min_{A,B} \frac{1}{N} \sum_{i=1}^{N} \log\left(1+\exp\left(-y_n\left(A\left(w_{SVM}^{T}\Phi(x)+b_{SVM}\right)+B\right)\right)\right)$$

- Here $A > 0$ and B~0, otherwise the solution of SVM is very BAD.
- This SVM is called “Probability SVM”,
- It was proposed by platt, so it’s called platt’s model. It runs SVM first then does logistic regression. But it is just an approximated solutions. Next we want to find a solution by logistic regression.

Key of the kernel trick

- Let's recall why does the kernel work, the optimal $\boldsymbol{w}_{opt}$ was combined by z linearly,

$$w_{opt} = \sum_{i=1}^{N} \beta_n z_n \rightarrow w_{opt}{}^T z_n = \sum_{i=1}^{N} \beta_n z_n{}^T z_n = \sum_{i=1}^{N} \beta_n K(x_n, x)$$

- The methods we've introduced are the same, so our goal is to represent $\boldsymbol{w}_{opt}$ by z.

SVM

$$\mathbf{w}_{\text{SVM}} = \sum_{n=1}^{N} (\alpha_n y_n)\mathbf{z}_n$$

α_n from **dual solutions**

PLA

$$\mathbf{w}_{\text{PLA}} = \sum_{n=1}^{N} (\alpha_n y_n)\mathbf{z}_n$$

α_n by **# mistake corrections**

LogReg by SGD

$$\mathbf{w}_{\text{LOGREG}} = \sum_{n=1}^{N} (\alpha_n y_n)\mathbf{z}_n$$

α_n by **total SGD moves**

Hsuan Tien Lin, mltech/205_handout

Representation theorem

- Claim: for any L2 regularized linear model

$$\boxed{\min_{w}\left(\frac{\lambda}{N}ww^T+\frac{1}{N}\sum_i err_i\right)} \qquad \boxed{w_{opt}=\sum_{i=1}^{N}\beta_i z_i}$$

- *Proof* :
- Let $\boldsymbol{w}_{opt} = \boldsymbol{w}_{\|} + \boldsymbol{w}_{\perp}$, $\boldsymbol{w}_{\|}$ is span by z_n, $\boldsymbol{w}_{\perp}$ and is linearly dependent to z_n. So That err(y_n, $\boldsymbol{w}_{opt}^{\boldsymbol{T}}$, z_n) = err(y_n, $(\boldsymbol{w}_{\|} + \boldsymbol{w}_{\perp})^{\boldsymbol{T}}$, z_n), then
- $\boldsymbol{w}_{opt}\, \boldsymbol{w}_{opt}^{\boldsymbol{T}} = \boldsymbol{w}_{\|}\boldsymbol{w}_{\|}^{\boldsymbol{T}} + \boldsymbol{w}_{\perp}\boldsymbol{w}_{\perp}^{\boldsymbol{T}} + 2\, \boldsymbol{w}_{\|}\boldsymbol{w}_{\perp} > \boldsymbol{w}_{\|}\boldsymbol{w}_{\|}^{\boldsymbol{T}}$ **(→←)**
- So $\boldsymbol{w}_{\perp}^{\boldsymbol{T}} = 0$

L2-regularized logistic regression

- So by the representation theorem, $w_{opt} = \sum_{i=1}^{N} \beta_i z_i$
- $\min_w \left(\frac{\lambda}{N} ww^T + \frac{1}{N} \sum_i err_i \right) = \min_w \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \beta_i \beta_j K(x_i, x_j) + \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + \exp \left(-y_n \sum_{j=1}^{N} \beta_j K(x_i, x_j) \right) \right) \right)$
- Here $K(x_i, x_j)$ is the kernel, $\sum_{j=1}^{N} \beta_j K(x_i, x_j)$ is the linear model,
- $\sum_{i=1}^{N} \sum_{j=1}^{N} \beta_i \beta_j K(x_i, x_j) = \beta K \beta^T$ is the regularizer

L2-regularized logistic regression

- So it can be seen as a linear model of β_i with kernel as transformation and kernel regularized. It's like SVM
- The β_i is not often 0 but α_i in SVM is often 0