In-Class 7

Wednesday, December 4, 2019 2:29 PM

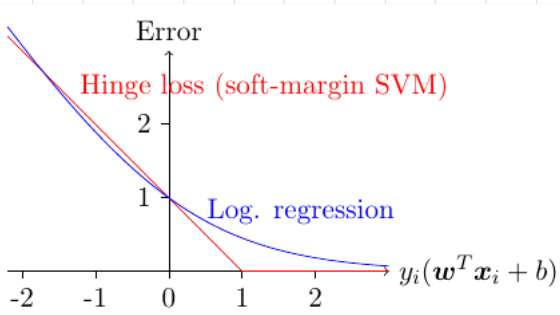
Problem 7: What is the connection between soft-margin SVM and logistic regression?

SVM

$$\min_{w,b} \frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i$$

$$y_i (w^T x_i + b) - 1 + \xi_i \geq 0$$

$$\xi_i \geq 0$$



$$\text{hinge}(z) = \max(0, 1 - z)$$

hinge loss formulation

$$\min_{w,b} E(w, b, C) = \frac{1}{2C} w^T w + \sum_{i=1}^{N} \underbrace{\text{hinge}(y_i (w^T x_i + b))}_{\text{loss}}$$

Logistic Regression:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$$p(y_i = 1 \mid x_i, (w, b)) = \sigma(w^T x_i + b)$$

$$p(y_i = -1 \mid x_i, (w, b)) = 1 - \sigma(w^T x_i + b) = \sigma(-(w^T x_i + b))$$

$$\rightarrow p(y_i \mid x_i, (w, b)) = \sigma(y_i (w^T x_i + b))$$

Altogether we get $p(y \mid X, (w, b)) = \prod_{i=1}^{N} \sigma(y_i (w^T x_i + b))$

$$\min_{w,b} E(w, b) = -\ln(p(y \mid X, (w, b))) = -\sum_{i=1}^{N} \ln((1 + e^{-y_i (w^T x_i + b)})^{-1})$$

$$= \sum_{i=1}^{N} \ln(1 + e^{-y_i (w^T x_i + b)})$$

add l_2-regularizer

$$\min_{w,b} E(w, b, \lambda) = \lambda w^T w + \sum_{i=1}^{N} \underbrace{\ln(1 + e^{-y_i (w^T x_i + b)})}_{\text{loss}}$$

Problem 8: Consider a soft-margin SVM fitted to a linearly separable dataset $\mathcal{D}$ using the Hinge loss formulation of the optimization task.

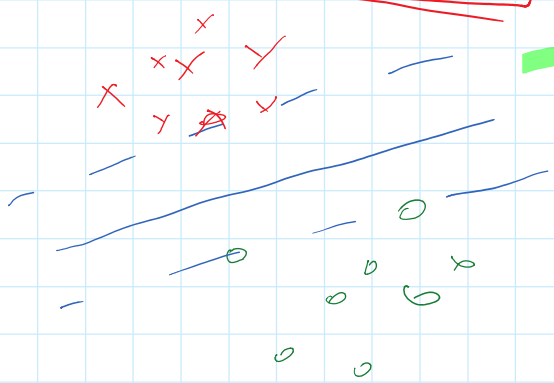
$$\text{minimize}_{\boldsymbol{w},b} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\max\{0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\}$$

a) Is it guaranteed that all training samples in $\mathcal{D}$ will be assigned the correct label by the model?

b) Prove that if for some $C_0 \geq 0$ the resulting model classifies all training samples correctly then it will also be the case if we train the model with any larger $C \geq C_0$.

b.) Denote by $h(w,b) = \sum_{i=1}^{N} \max(0, 1 - y_i(w^T x_i + b))$ and let $C > C_0 \geq 0$. Note that all points lie outside of the margin if and only if $h(w,b) = 0$.

and all samples lie outside of the margin



$$(w^*, b^*) = \underset{w,b}{\operatorname{argmin}} \; \frac{1}{2} w^T w + C_0 h(w,b)$$

$$(v^*, d^*) = \underset{w,b}{\operatorname{argmin}} \; \frac{1}{2} w^T w + C h(w,b)$$

$$\Rightarrow \quad \frac{1}{2}\|w^*\|^2 + C_0 h(w^*, b^*) \leq \frac{1}{2}\|v^*\|^2 + C_0 h(v^*, d^*)$$

$$\frac{1}{2}\|v^*\|^2 + C h(v^*, d^*) \leq \frac{1}{2}\|w^*\|^2 + C h(w^*, b^*)$$

Sum the two inequalities up

$$\Rightarrow \frac{1}{2}\|w^*\|^2 + \frac{1}{2}\|v^*\|^2 + C_0 h(w^*, b^*) + C h(v^*, d^*) \leq$$

$$\frac{1}{2}\|v^*\|^2 + \frac{1}{2}\|w^*\|^2 + C_0 h(v^*, d^*) + C h(w^*, b^*)$$

$$\Rightarrow (C - C_0)\left(h(v^*, d^*) - h(w^*, b^*)\right) \leq 0$$

≥ 0

$\Rightarrow \quad h(v^*, d^*) \leq h(w^*, b^*) = 0$

$\Rightarrow \quad h(v^*, d^*) = 0 \quad \Rightarrow 1 - y_i (v^{*T} x_i + d^*) \leq 0$

In-Class 10a

Wednesday, December 4, 2019 2:31 PM

Problem 10: Consider the Gaussian kernel

$$k_G(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2}{2\sigma^2}\right), \text{ with } \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d.$$

a) Suppose you have found a feature map $\theta : \mathbb{R}^d \to \mathbb{R}^{d_1}$ that transforms your data into a feature space in which a SVM with a Gaussian kernel works well. However, computing $\theta(\boldsymbol{x})$ is computationally expensive and luckily you discover an efficient method to compute the scalar product

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \theta(\boldsymbol{x}_1)^T \theta(\boldsymbol{x}_2)$$

in your feature space without having to compute $\theta(\boldsymbol{x}_1)$ and $\theta(\boldsymbol{x}_2)$ explicitly. Show how you can use the scalar product $k(\boldsymbol{x}_1, \boldsymbol{x}_2)$ to efficiently compute the Gaussian kernel $k_G(\theta(\boldsymbol{x}_1), \theta(\boldsymbol{x}_2))$ in your feature space.

$$k_G(\theta(x_1), \theta(x_2)) = \exp\left(-\frac{\|\theta(x_1) - \theta(x_2)\|^2}{2\sigma^2}\right)$$

$$\|a-b\|^2 = (a-b)^T(a-b) = a^Ta - 2a^Tb + b^Tb$$

$$= \exp\left(-\frac{\theta(x_1)^T\theta(x_1) - 2\theta(x_1)^T\theta(x_2) + \theta(x_2)^T\theta(x_2)}{2\sigma^2}\right) =$$

$$= \exp\left(-\frac{k(x_1,x_1) - 2k(x_1,x_2) + k(x_2,x_2)}{2\sigma^2}\right)$$

Problem 10: Consider the Gaussian kernel

$$k_G(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2}{2\sigma^2}\right), \text{ with } \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d.$$

b) One of the nice things about kernels is that new kernels can be constructed out of already given ones. Use the five kernel construction rules from the lecture to prove that k_G is a kernel.

Hint: Use the Taylor expansion of the exponential function to prove that $\exp \circ k_1$ *is a kernel if* k_1 *is a kernel. Also, consider* $k_2(\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2))$ *with the linear kernel* $k_2(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1^T \boldsymbol{x}_2$ *and a feature map* ϕ *with only one feature.*

Let $k_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be kernels, with $\mathcal{X} \subseteq \mathbb{R}^N$. Then the following functions k are kernels as well:

- $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = k_1(\boldsymbol{x}_1, \boldsymbol{x}_2) + k_2(\boldsymbol{x}_1, \boldsymbol{x}_2)$ (1)
- $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = c \cdot k_1(\boldsymbol{x}_1, \boldsymbol{x}_2)$, with $c > 0$ (2)
- $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = k_1(\boldsymbol{x}_1, \boldsymbol{x}_2) \cdot k_2(\boldsymbol{x}_1, \boldsymbol{x}_2)$ (3)
- $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = k_3(\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2))$, with the kernel k_3 on $\mathcal{X}' \subseteq \mathbb{R}^M$ and $\phi : \mathcal{X} \to \mathcal{X}'$ (4)
- $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{x}_1^T \boldsymbol{A} \boldsymbol{x}_2$, with $\boldsymbol{A} \in \mathbb{R}^N \times \mathbb{R}^N$ symmetric and positive semidefinite (5)
- $k_i(x_1, x_2)$ are valid kernels and $k(x_1, x_2) = \lim_{i \to \infty} k_i(x_1, x_2)$ $\forall x_1, x_2$ is also a valid kernel. (6)

First we show that $\exp(k(x_1, x_2))$ is a kernel if $k(x_1, x_2)$ is a kernel.

$$\exp(k(x_1, x_2)) = 1 + \sum_{n=1}^{\infty} \frac{1}{n!} k(x_1, x_2)^n =$$

$$= 1 + \lim_{N \to \infty} \sum_{n=1}^{N} \frac{1}{n!} k(x_1, x_2)^n$$

(3) $\longrightarrow$ $k(x_1, x_2) \cdot k(x_1, x_2) = k(x_1, x_2)^2$ is a kernel

$k(x_1, x_2)^n$ is a kernel

(2) $\longrightarrow$ $\frac{1}{n!} k(x_1, x_2)^n$ is a kernel

(1) $\longrightarrow$ $\sum_{n=1}^{N} \frac{1}{n!} k(x_1, x_2)^n$ is a kernel

(1) $\rightarrow \sum_{n=1}^{\infty} \frac{1}{n!} k(x_1, x_2)$ is a kernel

(6) $\rightarrow \sum_{n=1}^{\infty} \frac{1}{n!} k(x_1, x_2)^n$ — // —

(1) $\rightarrow 1 + \sum_{n=1}^{\infty} \frac{1}{n!} k(x_1, x_2)^n = \exp(k(x_1, x_2))$ is a kernel ✓

a) $\rightarrow k_4(x_1, x_2) = \exp\left(-\frac{x_1^T x_1}{2\sigma^2}\right) \exp\left(-\frac{x_2^T x_2}{2\sigma^2}\right) \exp\left(\frac{x_1^T x_2}{\sigma^2}\right)$

(2), (5) $\rightarrow k(x_1, x_2) = x_1^T x_2 \frac{1}{\sigma^2}$ is a kernel

$\Rightarrow \exp\left(\frac{1}{\sigma^2} x_1^T x_2\right)$ is a kernel

(4) $\rightarrow k(x_1, x_2) = k_3(\phi(x_1), \phi(x_2))$ is a kernel, da $k_3(x_1, x_2) = x_1 \cdot x_2$ und $\phi(x) = \exp\left(-\frac{x^T x}{2\sigma^2}\right)$

(3) $\rightarrow k_4(x_1, x_2)$ is a kernel

Problem 10: Consider the Gaussian kernel

$$k_{\mathrm{G}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\frac{\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2}{2\sigma^2}\right), \text{ with } \boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d.$$

c) Can any finite set of points be linearly separated in the feature space of the Gaussian kernel if σ can be chosen freely?

Consider the limit $\sigma \to 0$.

$$k_G(x_1, x_2; \sigma) \xrightarrow{\sigma \to 0} k(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 = x_2 \\ 0 & \text{if } x_1 \neq x_2 \end{cases}$$

All training samples are correctly classified if

$$y_i\left(w^T \phi_G(x_i) + b\right) > 0 \quad \forall i$$

$$k_G(x_1, x_2) = \phi_G(x_1)^T \phi_G(x_2)$$

Substitute $w = \sum_j y_j \alpha_j \phi_G(x_j)$ into

$$y_i\left(\sum_j y_j \alpha_j \underbrace{\phi_G(x_j)^T \phi_G(x_i)}_{k_G(x_j, x_i; \sigma)} + b\right) > 0$$

$$\xrightarrow{\sigma \to 0} \underbrace{y_i^2}_{1} \alpha_i + y_i b = \alpha_i + y_i b$$

By choosing $b = 0$ and $\alpha_i > 0$ we can fulfill if σ small.