RWTH AACHEN UNIVERSITY

6.5. The SMO Algorithm

Sequential Minimal Optimization

- Select a pair (i,j) of λ's to update
- Optimize the obj. fct. w.r.t. (λ_i, λ_j)
- Explicit solution ~~for~~ $(\lambda_i^*, \lambda_j^*)$
- Iterate until convergence

$\lrcorner$

6.6. Kernels

Instead of applying SVM to the raw data ("attributes") x_i apply it to transformed data ("features") $\phi(x_i)$.

ϕ is called feature mapping.

Aim: achieve better separability.

$$(D) \qquad \max_{\lambda} g(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \lambda_i \lambda_j x_i^T x_j$$

$g(\lambda)$ only depends on the inner product $x_i^T x_j$.

Substitute x_i by $\phi(x_i)$ and use some inner product $\langle \cdot | \cdot \rangle$.

Replace $x_i^T x_j$ by $K(x_i, x_j) = \langle \phi(x_i) | \phi(x_j) \rangle$

Remark: $K(x,y)$ is often easier to compute than $\phi(x)$ itself.

Thinking the Future
Zukunft denken

Intuition:

If $\phi(x), \phi(y)$ are close $\langle \phi(x), \phi(y) \rangle$ is large.

If $\phi(x) \perp \phi(y)$ then $\langle \phi(x), \phi(y) \rangle = 0$. Hence, $K(x,y)$ measures how similar x and y are.

Needed: an inner product in some feature space

$$\{\phi(x) \mid x \in \mathbb{R}^p\}.$$

<u>Example. 6.2.</u>

$x, y \in \mathbb{R}^p$, $K(x,y) = \langle x,y \rangle^2 = \left(\sum_{i=1}^{p} x_i y_i\right)^2$

Question: Is there some ϕ such that $\langle x,y \rangle^2$ is an inner product in the feature space.

$p=2$: $x = (x_1, x_2)^T$, $y = (y_1, y_2)^T$

Use $\phi(x) = (x_1^2, x_2^2, x_1 x_2, x_2 \cdot x_1) : \mathbb{R}^2 \to \mathbb{R}^4$

$$\begin{aligned} \langle \phi(x), \phi(y) \rangle &= x_1^2 y_1^2 + x_2^2 y_2^2 + x_1 x_2 y_1 y_2 + x_2 x_1 y_2 y_1 \\ &= x_1^2 y_1^2 + x_2^2 y_2^2 + 2 x_1 x_2 y_1 y_2 \\ &= (x_1 y_1 + x_2 y_2)^2 = \langle x,y \rangle^2. \quad \lrcorner \end{aligned}$$

<u>Example 6.3.</u> (Gaussian kernel)

$x, y \in \mathbb{R}^p$, $K(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$

Question: $\exists$ feature mapping ϕ and a feature space with $\langle \cdot, \cdot \rangle$?

<u>Def. 6.4.</u> Kernel $K(x,y)$ is called <u>valid</u> if there exists a feature function ϕ such that

$$K(x,y) = \langle \phi(x), \phi(y) \rangle \text{ for all } x,y \in \mathbb{R}^p. \quad \lrcorner$$

<u>Theorem 6.5.</u> (Mercer)

Given $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. K is a valid kernel if and only if for any $x_1, \dots, x_n \in \mathbb{R}^p$ the kernel matrix

$$\big(K(x_i,x_j)\big)_{i,j=1,\dots,n} \text{ is n.n.d.} \quad \lrcorner$$

<u>Proof.</u> only "$\Rightarrow$":

$$K \text{ valid} \Rightarrow \exists\, \phi : K(x_i,x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \langle \phi(x_j), \phi(x_i) \rangle$$

hence symmetry. Moreover

$$
\begin{aligned}
& z^T \big(K(x_i,x_j)\big)_{i,j=1,\dots,n} z \\
&= z^T \big(\langle \phi(x_i), \phi(x_j) \rangle\big)_{i,j} z \\
&= \sum_{k,\ell} z_k z_\ell \langle \phi(x_k), \cancel{x_\ell)} \phi(x_\ell) \rangle \\
&= \Big\langle \sum_k z_k \phi(x_k), \sum_\ell z_\ell \phi(x_\ell) \Big\rangle \geq 0. \quad \boxtimes
\end{aligned}
$$

Example 6.6. (Polynomial kernel)

$K(x,y) = (x^T y + c)^d$, $x, y \in \mathbb{R}^p$, $c \in \mathbb{R}$, $d \in \mathbb{N}$, $d \geq 2$

Feature space of dim $\binom{p+d}{d}$ containing all monomials of degree $\leq d$. ┘ (Ex) determine ϕ

7. Machine Learning

7.1 Supervised Learning

Given (x_i, y_i), $i = 1, \dots, n$, training examples/samples.

$x_i \in \mathcal{X}$: input variables, feature variables

$y_i \in \mathcal{Y}$: output variables, target variables

$\{(x_i, y_i) \mid i = 1, \dots, n\}$ is called training set.

Supervised learning problem: determine a function $h : \mathcal{X} \to \mathcal{Y}$ (a "hypothesis")

so that $h(x)$ is a "good" predictor of y.

If $\mathcal{Y}$ is continuous : regression problem

$\mathcal{Y}$ is discrete : classification problem

7.1.1 Linear Regression

Training examples $\underline{x}_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$, $i = 1, \dots, n$

$$y_i = \vartheta_0 + x_{i1}\vartheta_1 + \dots + x_{ip}\vartheta_p + \varepsilon_i \,, \quad \varepsilon_i \text{ random error}$$

$$= (1, \underline{x}_i^T)\,\underline{\vartheta} + \varepsilon_i$$

Hence, learn $h_\vartheta(\underline{x}) = (1, \underline{x}^T)\vartheta$, $\vartheta = (\vartheta_0, \vartheta_1, \dots, \vartheta_p)^T$ parameter

Set $\overset{n \times (p+1)}{X} = \begin{pmatrix} 1 & \underline{x}_1^T \\ \vdots & \vdots \\ 1 & \underline{x}_n^T \end{pmatrix}$, $\underline{y} = (y_1, \dots, y_n)^T$, $\underline{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$

Then $\underline{y} = X\underline{\vartheta} + \underline{\varepsilon}$

Problem: find the best $\underline{\vartheta}$ by solving

$$\min_{\vartheta \in \mathbb{R}^{p+1}} \| \underline{y} - X\underline{\vartheta} \|$$

Solution (i) project y onto $\operatorname{Im}(X)$: $\hat{y}$

(ii) find ϑ s.t. $\hat{y} = X\vartheta$

(i) $X(X^TX)^{-1}X^T$ is an orth. proj. onto $\operatorname{Im}(X)$ provided $(X^TX)^{-1}$ exists.

$$\hat{y} = X(X^TX)^{-1}X^Ty = \arg\min_{z \in \operatorname{Im}(X)} \|y - z\|$$

(ii) $\hat{y} = X\vartheta \Rightarrow \cancel{X^T} X \cancel{(X^TX)^{-1}} X^T y = X^T X \vartheta$

$\cancel{\Rightarrow \vartheta^* = (X^TX)^{-1} X^T y}$

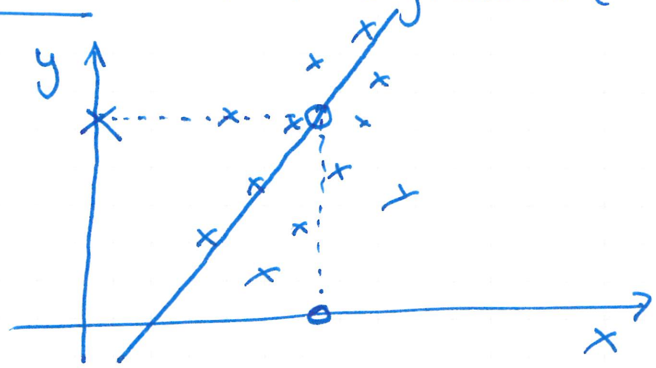
$\Rightarrow \vartheta^* = (X^TX)^{-1}X^Ty$

is a solution.

In summary $\vartheta^* = (X^T X)^{-1} X^T y$ is a solution.

Note: the inverse $(X^T X)^{-1}$ must exist. If not, replace $(X^T X)^{-1}$ by the so called Moore-Penrose inverse $(X^T X)^+$.

Example. Linear Regression (1-dim)



Solution

$$\vartheta_1^* = \frac{\sigma_{xy}}{\sigma_x^2}, \quad \vartheta_0^* = \bar{y} - \vartheta_1^* \bar{x}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\sigma_{xy} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}, \quad \sigma_x^2 = \sigma_{xx}$$