## 7.1.2. Logistic Regression

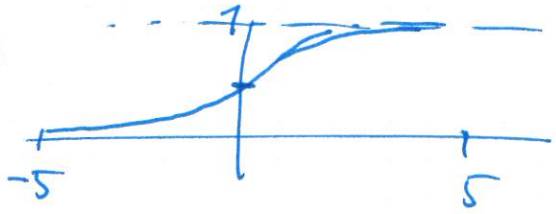
Classification like regression with finitely many values of $y$. In the following only $y \in \{0, 1\}$.

Hypothesis:

$$h_\vartheta(x) = g(\vartheta^T x) = \frac{1}{1+e^{-\vartheta^T x}} \in [0,1]$$

$g(z) = \frac{1}{1+e^{-z}}$ is called logistic or sigmoid fct.

Plot:




It holds $g'(z) = g(z)\,(1-g(z))$

Given training $(x_i, y_i)$, $i = 1, \dots, m$

$x_i \in \mathbb{R}^p$, $y_i \in \{0,1\}$

As before set $x_{i0} = 1$, s.t.

$$\vartheta^T x_i = \vartheta_0 + \sum_{j=1}^{p} \vartheta_j x_{ij}, \quad x_i = (1, x_{i1}, \dots, x_{ip})$$

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

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

(ii) $\hat{y} = X\beta$

$\Rightarrow$ ~~$X^TX(X^TX)^{-1}X^Ty = X^TX\beta$~~

$\Rightarrow X^Ty = X^TX\beta$ (normal equations)

~~and $\beta^*$~~

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

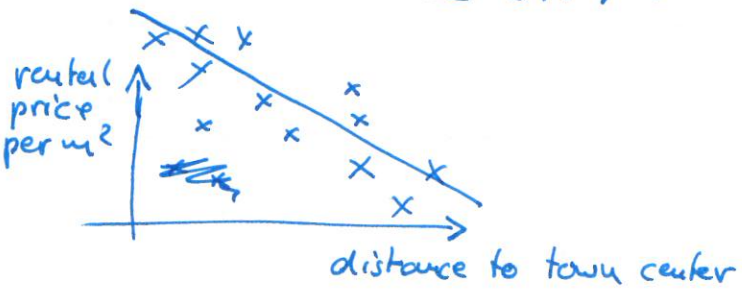
and $X\beta^* = X(X^TX)^{-1}X^Ty = \hat{y}$

In summary:

$\beta^* = (X^TX)^{-1}X^Ty$ is a solution of (1).

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

Example.



## 7.1.1. Linear Regression

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

Assumption / Hypothesis

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

$\varepsilon_i$: random vector

$$= (1, x_i^T)\vartheta + \varepsilon_i$$

Hence, learn $h_\vartheta(x) = (1, x^T)\vartheta$,

$\vartheta = (\vartheta_0, \vartheta_1, \dots, \vartheta_p)^T$ parameter

Set $X^{n\times(p+1)} = \begin{pmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_n^T \end{pmatrix}$, $y = (y_1, \dots, y_n)^T$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$

Then $y = X\vartheta + \varepsilon$ ($n$ rows, $n\times(p+1)$, $p+1$, $n$)

Problem: Find the "best" $\vartheta$

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

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

(ii) find $\vartheta$ such that $\hat{y} = X\vartheta$

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 $\leq d$. ⌋

# 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

$$h : \mathcal{X} \to \mathcal{Y}$$

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

If $y$ continuous: regression problem

$y$ discrete: classification problem

**Def. 6.4.** Kernel $K(x,y)$ is called valid if there is a feature function $\phi$ s.t.

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

**Th. 6.5** (Mercer)

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

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

**Proof.** only "$\Rightarrow$"

$K$ 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

$$z^T \big(K(x_i, x_j)\big)_{i,j=1,\dots,n} z$$

$$= z^T \big(\langle \phi(x_i), \phi(x_j) \rangle\big)_{i,j=1,\dots,n} z = \sum_{k,\ell} z_k z_\ell \langle \phi(x_k), \phi(x_\ell) \rangle$$

$$= \Big\langle \sum_k z_k \phi(x_k), \sum_\ell z_\ell \phi(x_\ell) \Big\rangle \geq 0 \qquad \blacksquare$$

Needed: inner product in the feature space $\{\phi(x) \mid x \in \mathbb{R}^p\}$.

Example. 6.2.

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

Question: Is there some $\phi$ s.t. $\langle x,y \rangle^2$ is an inner product in the feature space?

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

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

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

□

Example 6.3. (Gaussian kernel)

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

Question: $\exists$ feature fct. $\phi$ and a feature space with $\langle \cdot, \cdot \rangle$?

## 6.6. Kernels

raw data (attributes)

$\longrightarrow$ transformed data (features)

$x_i \longrightarrow \phi(x_i)$, $\phi$ feature mapping

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

s.t. $0 \leq \lambda \leq C$

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

$g(\lambda)$ only depends on the inner products

$$x_i^T x_j = \langle x_i, x_j \rangle$$

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.

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

$\phi(x) \perp \phi(y)$ then $\langle \phi(x), \phi(y) \rangle = 0$. Hence, $K(x,y)$ measures similarity of $x$ and $y$.