Given two collections of points A and B.

We wish to find (ω, b) s.t. $\quad \omega^T\phi(x)+b > 0$ if $x\in A$

$\omega^T\phi(x)+b < 0$ if $x\in B$.

This problem is formulated as:

$$(P)\quad \begin{cases} \min\limits_{w\in\mathbb{R}^k,\, b\in\mathbb{R}} & \|\omega\|_2^2 \\ \text{s.t.} & 1-y^i(\omega^T\phi(x^i)+b)\le 0, \\ & i=1,2,\dots N. \end{cases}$$

[$\phi: \mathbb{R}^n \to \mathbb{R}^K$ maps points $x\in\mathbb{R}^n$ to features $\phi(x)\in\mathbb{R}^K$

N: # labeled data that we have $(y^1, x^1), (y^2, x^2), \dots$

$y^i = 1 \iff x^i\in A$.

$y^i = -1 \iff x^i \in B$

Dual of (P) is given by:

$$\begin{aligned} \min_{\lambda\in\mathbb{R}^N} \quad & -\sum_{i=1}^N \lambda_i + \frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N \lambda_i\lambda_j\, y^i y^j\, \phi(x^i)^T\phi(x^j) \\ \text{s.t.} \quad & \sum_{i=1}^N \lambda_i y^i = 0, \quad \lambda \ge 0 \end{aligned}$$

$= K(x^i, x^j)$

Once the optimal solution (ω^*, b^*) is obtained, then for a new point $\bar{x}$, we can determine its label as:

$$\bar{y} = \text{sign}\left[(\omega^*)^T\phi(\bar{x}) + b^*\right] \qquad \omega^* = \sum_{i=1}^N \lambda_i^* y^i \phi(x^i)$$

+ve $\Rightarrow \bar{x}\in A$ $\qquad (\omega^*)^T\phi(\bar{x}) = \sum_{i=1}^N \lambda_i^* y^i \phi(x^i)^T\phi(\bar{x})$

−ve $\Rightarrow \bar{x}\in B$

Choice of ϕ is not unique, and can vary significantly.

One systematic way of dealing with it is in terms of Kernel functions.

A function $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called a Kernel function if

i) $K(x^i, x^j) = K(x^j, x^i)$ (symmetric)

ii) positive semidefinite property:

given $(x^1, x^2, \ldots x^N)$, we can create a matrix

$\bar{K} \in \mathbb{R}^{N \times N}$, $[\bar{K}]_{ij} = K(x^i, x^j)$

$$\bar{K} = \begin{bmatrix} K(x^1, x^1) & K(x^1, x^2) & \cdots \\ K(x^2, x^1) & K(x^2, x^2) & \cdots \\ \vdots & & \vdots \end{bmatrix}$$

$\bar{K}$ is positive semidefinite.

From the theory of reproducing kernel Hilbert spaces, we have the property that every kernel function K has an associated feature map:

$$\phi(x) = K(x, \cdot) \qquad \phi : \mathbb{R}^n \to \mathcal{C}$$

$\uparrow$ class of functions.

$$\phi(x^i)^T \phi(x^j) = K(x^i, x^j) \in \mathbb{R}.$$

Examples of Kernel

i) Gaussian: $K(x_i, x_j) = e^{-c\|x_i - x_j\|_2^2}$

ii) polynomial: $K(x_i, x_j) = (x_i^T x_j + c)^p$

The nonlinear classification problem can now be written as

$$\min_{\lambda \in \mathbb{R}^N} \; -\sum_{i=1}^{N} \lambda_i + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y^i y^j \underline{K(x^i, x^j)}$$

$$\text{s.t.} \quad \lambda \geq 0, \quad \sum_{i=1}^{N} \lambda_i y^i = 0$$

Once we solve for λ^*, we find

$$w^* = \sum_{i=1}^{N} \lambda_i^* y^i \phi(x^i)$$

pick index i s.t. $\lambda_i > 0$.

$$b^* = \frac{1}{y^i} - (w^*)^T \phi(x^i)$$

$$= \frac{1}{y^i} - \sum_{j=1}^{N} \lambda_j^* y^j K(x^j, x^i)$$

$b^* \in \mathbb{R}$ using complementary slackness.

If we encounter a new point $\bar{x}$, evaluate

$$(w^*)^T \phi(\bar{x}) + b^* = \underline{\sum_{i=1}^{N} \lambda_i^* y^i \underline{K(x^i, \bar{x})} + b^*}$$

if +ve, $\bar{x} \in A$

−ve, $\bar{x} \in B$.

Regression problems

Given N input-output pairs $(x^1, y^1), (x^2, y^2), \ldots (x^N, y^N)$,

$x^i \in \mathbb{R}^n$, $y^i \in \mathbb{R}$.

We wish to learn a function $f(x) \approx y$.

Suppose we hypothesize the function $f(x) = w^T \phi(x)$,

where ϕ is a feature map.

$\phi : \mathbb{R}^n \to \mathbb{R}^K$.

for a given $\bar{w}$

Residual error at point i is $\underline{(\bar{w})^T \phi(x^i) - y^i}$

we now formulate the problem

$$\min_{w \in \mathbb{R}^K} \underbrace{\sum_{i=1}^{N} \left(w^T \phi(x^i) - y^i\right)^2}$$

$$\boxed{\min_{w \in \mathbb{R}^K} \underline{\|\phi(x) w - y\|_2^2}},$$

"least squares problem"

where $y \in \mathbb{R}^N = \begin{bmatrix} y^1 \\ y^2 \\ \vdots \\ y^N \end{bmatrix}$

$$\underline{\phi(x) \in \mathbb{R}^{N \times K}} = \begin{bmatrix} - \phi(x^1)^T - \\ - \phi(x^2)^T - \\ \vdots \\ - \phi(x^N)^T - \end{bmatrix}, \quad \phi(x^i) \in \mathbb{R}^K$$

Q Is this problem always convex? yes.

cost function $g(w) = \|\phi(x) w - y\|_2^2 = (\phi(x)w - y)^T(\phi(x)w - y)$

$$= w^T \phi(x)^T \phi(x) w - 2 y^T \phi(x) w + y^T y$$

$\nabla_w^2 g(w) = 2\phi(x)^T \phi(x)$ always positive semidefinite.

$$\nabla g(w) = 0 \Rightarrow \underline{\phi(x)^T \phi(x)} w - y^T \phi(x) = 0$$

↳ may not be invertible in which case sol^n may not be unique.

In addition to minimizing 2-norm of residual error, one may minimize 1-norm & ∞-norm.

(P1) $\min_{w \in \mathbb{R}^K} \|\phi(x) w - y\|_1$

(P2) $\min_{w \in \mathbb{R}^K} \underline{\|\phi(x) w - y\|_\infty}$

Both of these are linear programs

$$(P2) \quad \min_{w \in \mathbb{R}^K} \max_{1 \le i \le N} |\phi(x^i)^T w - y^i|$$

$$\Updownarrow$$

$$\min_{w \in \mathbb{R}^K, t \in \mathbb{R}} t$$

$$\text{s.t.} \quad t = \max_{1 \le i \le N} |\phi(x^i)^T w - y^i|$$

$$\Updownarrow$$

$$|\phi(x^i)^T w - y^i| = \max(\phi(x^i)^T w - y^i, -\phi(x^i)^T w + y^i)$$

$$t \ge |\phi(x^i)^T w - y^i|, \quad i = 1, 2, \ldots N$$

which is a linear program

$$\min_{w \in \mathbb{R}^K, t \in \mathbb{R}} t$$

$$\text{s.t.} \quad t \ge \phi(x^i)^T w - y^i, \quad i = 1, 2, \ldots N$$

$$t \ge -\phi(x^i)^T w + y^i, \quad i = 1, 2, \ldots N.$$

$$(P1) \quad \min_{w \in \mathbb{R}^K} \|\phi(x) w - y\|_1 = \min_{w \in \mathbb{R}^K} \sum_{i=1}^N \underbrace{|\phi(x^i)^T w - y^i|}_{=: t_i}$$

$$t_i = \max(\quad)$$

linear program

$$\min_{w \in \mathbb{R}^K, t \in \mathbb{R}^N} \sum_{i=1}^N t_i$$

$$\text{s.t.} \quad t_i \ge \phi(x^i)^T w - y^i, \quad i = 1, 2, \ldots N$$

$$t_i \ge -\phi(x^i)^T w + y^i, \quad i = 1, 2, \ldots N$$

6th March 2024

Least squares problem : $\min_{w \in \mathbb{R}^K} \underbrace{\|y - \phi(x) w\|_2^2}_{f(w)}$, $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$

$\phi \in \mathbb{R}^{N \times K}$

$$f(w) = (y - \phi(x) w)^T (y - \phi(x) w)$$
$$= y^T y + w^T \phi(x)^T \phi(x) w - 2 y^T \phi(x) w$$
$$\nabla_w f(w) = 2\phi(x)^T \phi(x) w - 2\phi(x)^T y = 0$$
$$\Rightarrow \boxed{\phi(x)^T \phi(x) w^* = \phi(x)^T y} \quad \text{--- (1)}$$

Is $\phi(x)^T \phi(x)$ always invertible ?

If $K > N$: $\text{rank}(\phi(x)) < K$

$\Downarrow$

$\phi(x)$ is a fat matrix

$\Rightarrow \phi(x)$ has a non-empty nullspace.

Let $\eta \in \text{Null}(\phi(x)) \Leftrightarrow \phi(x)\eta = 0$.

Then if w^* satisfies (1), then $w^* + \eta$ also satisfies (1)

- $\Rightarrow$ weights satisfying (1) is not unique,
- $\Rightarrow$ weights assigned to different features vary wildly.
- $\Rightarrow$ prediction on new data-points is not unique & varies wildly.

To tackle this issue, a regularizer term is added to the cost function.

$$\min_{w \in \mathbb{R}^K} \|y - \phi(x) w\|_2^2 + \lambda R(w).$$

$\lambda \geq 0$: hyper-parameter.

convex when $R(w)$ is convex.

Different choice of $R(\omega)$:

i) $R(\omega) = \|\omega\|_2^2$

ii) $R(\omega) = |\omega|_1 \Rightarrow$ optimal solution/weights has many entries as "0" (zero).

$$f(\omega) = \|y - \phi(x)\omega\|_2^2 + \lambda \omega^T \omega$$

$$\nabla f(\omega) = 2\big(\underbrace{\phi(x)^T\phi(x) + \lambda I}\big)\omega - 2\phi(x)^T y$$

positive definite, hence invertible.

Estimation Problems

$x \in \mathbb{R}^n$: quantity we are trying to estimate

$y \in \mathbb{R}$: observation.

Given $\hat{y}$, the maximum-likelihood estimate of x is defined as

$$\hat{x}_{ML}(\hat{y}) = \underset{x}{\operatorname{argmax}}\ f_Y(\hat{y}; x)$$

, where $f_Y(\hat{y}; x)$: density of y parametrized by x.

$$\Updownarrow$$

$$= \underset{x}{\operatorname{argmax}}\ \log(f_Y(\hat{y}; x))$$

$$\Updownarrow$$

$$= \boxed{\underset{x}{\operatorname{argmin}}\ -\log(f_Y(\hat{y}; x))}$$

Let $y = a^T x + v$, a: Known vector, $v \sim N(0, \sigma^2)$ Gaussian, zero-mean, variance σ^2.

observations : $(\hat{y}_1, \hat{y}_2, \dots \hat{y}_N)$, i.i.d

For $y = \hat{y}_1$, $f_Y(\hat{y}_1; x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left[\frac{\hat{y}_1 - a^T x}{\sigma}\right]^2\right)$

$y \sim N(a^T x, \sigma^2)$

$$\log\left(f_y(\hat{y}_1; x)\right) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{\hat{y}_1 - a^T x}{\sigma}\right)^2$$

$$f_Y(\hat{y}_1, \hat{y}_2 \cdots \hat{y}_N; x) = \prod_{i=1}^{N} f_Y(\hat{y}_1; x)$$

$$-\log(f_Y) = \sum_{i=1}^{N}\left(\frac{1}{2}\log(2\pi\sigma^2)\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(\hat{y}_i - a^T x)^2$$

Overall optimization problem:

$$\underset{x \in \mathbb{R}^n}{\operatorname{argmin}} -\log(f_Y) = \boxed{\underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^{N}(\hat{y}_i - a^T x)^2}$$

If the noise v has Laplace distribution: $f_V(v) = \frac{1}{\sqrt{2\alpha}} \exp\left(-\frac{|v|}{\alpha}\right)$

$$f_Y(\hat{y}_1; x) = f_V(\hat{y}_1 - a^T x) = \frac{1}{\sqrt{2\alpha}} \exp\left(-\left|\frac{\hat{y}_1 - a^T x}{\alpha}\right|\right)$$

If $y = \hat{y}_1$, $v = \hat{y}_1 - a^T x$

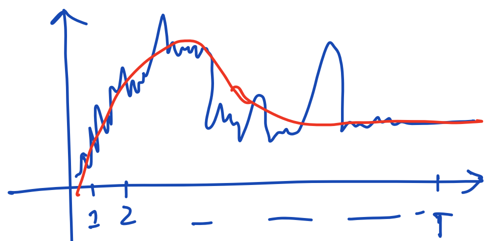
ML problem: $\underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^{N} |\hat{y}_i - a^T x|$: convex optimization problem.

Denoising :

x : true underlying signal

y : observation

$$y = x + w \quad , w : \text{noise}.$$



$(y_1, y_2 \cdots y_T)$

$$\min_{x} \underline{\|y - x\|_2^2} \Rightarrow \text{solution is } x^* = y.$$

If the signal is continuous, then consecutive elements of x will be close to each other.

Given: $y_1, y_2 \cdots y_T$

need to find: $(x_1, x_2 \cdots x_T) = x_{1:T}$

$$\min_{x_{1:T}} \sum_{i=1}^{N} (y_i - x_i)^2 + \lambda \underbrace{\sum_{j=1}^{N-1} (x_{j+1} - x_j)^2}_{\|Lx\|_2^2}, \quad L = \begin{bmatrix} -1 & 1 & - & - & \\ 0 & -1 & 1 & - & - \\ & - & - & - & \end{bmatrix}$$

$\Rightarrow$ **QP in $x_{1:T}$.**