Ch.4. Linear functions and Differentiation

§4.1 Linear functions.

Let $f: V \to \mathbb{R}$ be a function on a vector space V.

f is a linear function if

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \quad \forall\, \alpha, \beta \in \mathbb{R} \text{ and } x, y \in V.$$

Example 1: The mean of vectors in $\mathbb{R}^n$, i.e.,

$$f(x) = (x_1 + x_2 + \cdots + x_n)/n, \quad \text{for } x \in \mathbb{R}^n$$

is linear.

Example 2: The maximum element of vectors in $\mathbb{R}^n$, i.e.,

$$f(x) = \max\{x_1, x_2, \cdots, x_n\}, \quad \text{for } x \in \mathbb{R}^n$$

is NOT linear.

Example 3: $f: \mathbb{R}^n \to \mathbb{R}$ with $f(x) = a^T x$ is a linear function.

Example 4: $F: C[a,b] \to \mathbb{R}$ defined by

$F(f) = f(x)$, where $x \in [a,b]$ is a given number

is linear.

Example 5: $F: L^2(a,b) \to \mathbb{R}$ defined by

$F(f) = \int_a^b f(x)dx$ for $f \in L^2(a,b)$ is linear.

Example 6: $f: V \to \mathbb{R}$ (V is an inner product space) defined by

$f(x) = \langle x, z \rangle$, where $z \in V$ is a given vector in V.

Example 7: A norm function on a vector space V is NOT linear.

To see this,

$$\|-x\| = \|x\|,$$

which contradict with

$$f(-x) = f(-x + 0x) = -f(x) + 0f(x) = -f(x)$$

for f being linear on V.

Properties of linear functions

- Homogeneity: $f(\alpha x) = \alpha f(x) \quad \forall\, \alpha \in \mathbb{R}$ and $x \in V$.

  (Because $f(\alpha x) = f(\alpha x + 0x) = \alpha f(x) + 0\, f(x) = \alpha f(x)$)

  It implies $f(0) = 0$, because $f(0) = f(0 \cdot x) = 0 \cdot f(x) = 0 \quad \forall\, x \in V$.

- Additivity: $f(x+y) = f(x) + f(y) \quad \forall\, x, y \in V$.

- $f(\alpha_1 x_1 + \cdots + \alpha_k x_k) = \alpha_1 f(x_1) + \cdots + \alpha_k f(x_k), \quad \forall \alpha_1, \cdots, \alpha_k \in \mathbb{R},\ x_1, \cdots, x_k \in V$.

  To see this, we note that

$$\begin{aligned} f(\alpha_1 x_1 + \cdots + \alpha_k x_k) &= \alpha_1 f(x_1) + f(\alpha_2 x_2 + \cdots + \alpha_k x_k) \\ &= \alpha_1 f(x_1) + \alpha_2 f(x_2) + f(\alpha_3 x_3 + \cdots + \alpha_k x_k) \\ &\ \ \vdots \\ &= \alpha_1 f(x_1) + \cdots + \alpha_k f(x_k). \end{aligned}$$

Inner product representation of a linear function on Hilbert spaces

For simplicity, let's consider a linear function on $\mathbb{R}^n$ equipped with the standard inner product $\langle x, y \rangle = x^T y$ and the induced norm $\|x\|_2 = (\langle x, x \rangle)^{1/2}$.

- From the discussion above,

  For any given $a \in \mathbb{R}^n$, the function $f(x) = \langle a, x \rangle$ is linear.

- The reverse is true, i.e.,

  Any linear function $f : \mathbb{R}^n \to \mathbb{R}$ must be in the form of

  $f(x) = \langle a, x \rangle$ for some $a \in \mathbb{R}^n$.

To see this, let $e_1, e_2, \cdots, e_n$, where $e_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow$ ith component, be a basis of $\mathbb{R}^n$,

So that any $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$ is written as $x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$.

Therefore, if f is a linear function, then

$$f(x) = f(x_1e_1 + x_2e_2 + \cdots + x_ne_n)$$
$$= x_1 f(e_1) + x_2 f(e_2) + \cdots + x_n f(e_n) \quad \text{— by property of linear functions}$$
$$= \langle a, x \rangle,$$

where $a = \begin{pmatrix} f(e_1) \\ f(e_2) \\ \vdots \\ f(e_n) \end{pmatrix} \in \mathbb{R}^n$.

- Furthermore, the representation of a linear function $f(x) = \langle a, x \rangle$ is unique, which means there is only one vector $a \in \mathbb{R}$ for which $f(x) = \langle a, x \rangle$ holds for all x.

  Indeed, suppose that a is not unique, i.e., we have two vectors a, b such that $f(x) = \langle a, x \rangle$ and $f(x) = \langle b, x \rangle$ for all $x \in \mathbb{R}^n$.

  Then, let $x = e_i$: $f(e_i) = \langle a, e_i \rangle = a_i$ and $f(e_i) = \langle b, e_i \rangle = b_i$

  So, $a_i = b_i$, $i = 1, 2, \cdots, n$.

  Therefore $a = b$.

- Altogeter, we see that

  a linear function $f: \mathbb{R}^n \to \mathbb{R}$ if and only if $f(x) = \langle a, x \rangle$ for some unique $a \in \mathbb{R}^n$

The above holds true for any linear function on Hilbert spaces, widely known as Riesz representation theorem.

Theorem (Riesz representation thereom):

Let H be a Hilbert space, and f be a function: $H \to \mathbb{R}$. Then f is linear and bounded if and only if $f(x) = \langle a, x \rangle$ for some unique $a \in H$.

Example 1: We know mean(x) is linear on $\mathbb{R}^n$. Since $\mathbb{R}^n$ is a Hilbert space, we can find a unique $a \in \mathbb{R}^n$ s.t.

$$\text{mean}(x) = \langle a, x \rangle.$$

Indeed,

$$\text{mean}(x) = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}x_1 + \frac{1}{n}x_2 + \cdots + \frac{1}{n}x_n = \langle a, x \rangle,$$

where $a = \left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)^T$.

Example 2: Let H be a Hilbert space, and $\|\cdot\|$ is the norm.

It is known that the norm function is NOT linear.

Therefore,

there doesn't exist $a \in H$ such that $\|x\| = \langle a, x\rangle \quad \forall x \in H$.

Hyperplanes

- Again, we consider $\mathbb{R}^n$ as a Hilbert space, where any linear function is written as $\langle a, x\rangle$ for some $a \in \mathbb{R}^n$.

Consider the set

$$S_{a,0} = \{x \mid \langle a, x\rangle = 0\},$$

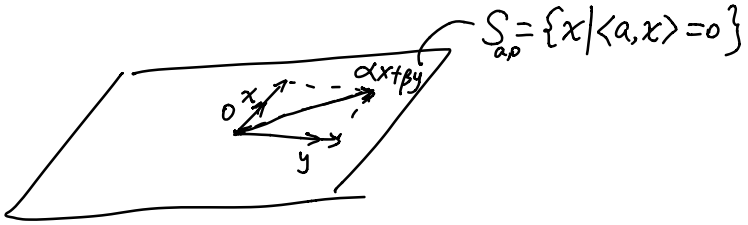
Then, $\forall\ x, y \in S_{a,0}$ and $\alpha, \beta \in \mathbb{R}$,

$$\langle a, \alpha x + \beta y\rangle = \alpha\langle a, x\rangle + \beta\langle a, y\rangle = 0 \Rightarrow \alpha x + \beta y \in S.$$

Therefore, $S_{a,0}$ is a plane.

Since the co-dimension of S_0 is 1 (because it is define by one linear equation)

$S_{a,0}$ is called a hyperplane.



Now let's consider

$$S_{a,b} = \{x \mid \langle a, x\rangle = b\} \quad \text{for } b \in \mathbb{R} \text{ is given.}$$

Let $x_0 \in S_{a,b}$, i.e., $\langle a, x_0\rangle = b$, be fixed.

Then $S_{a,b} = S_{a,0} + x_0$ because:

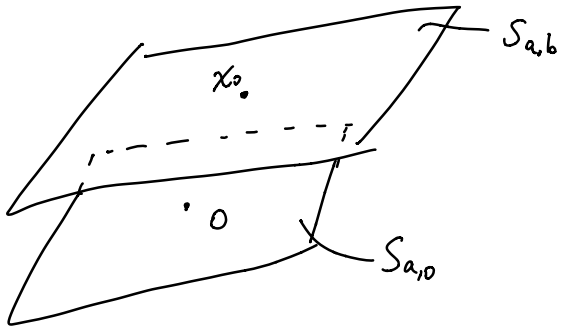
(1) $\forall\ x \in S_{a,b} \quad \langle a, x - x_0\rangle = \langle a, x\rangle - \langle a, x_0\rangle = 0, \Rightarrow x - x_0 \in S_{a,0}.$

$$\Rightarrow x \in S_{a,0} + x_0.$$

② $\forall x \in S_{a,0}$ $\quad \langle a, x+x_0 \rangle = \langle a, x \rangle + \langle a, x_0 \rangle = b \implies x+x_0 \in S_{a,b}$

In other words,

$S_{a,b}$ is a shift of a hyperplane, still called a hyperplane.



- This concept can be generalized to any inner product space V.

The set $\{x \in V \mid \langle a, x \rangle = b\}$, where $a \in V$ and $b \in \mathbb{R}$ are given, is called a Hyperplane in V.

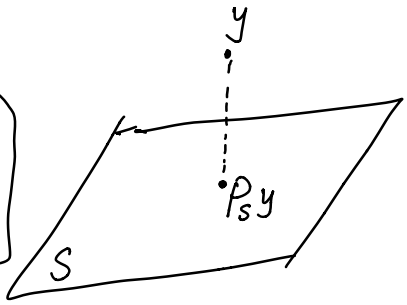
Projection onto hyperplanes

- Consider a Hilbert space V and a hyperplane S

$$S = \{x \in V \mid \langle a, x \rangle = b\}$$

Let $y \in V$ be a given vector.

The vector on S that is the closest to y is called the projection of y on S, denoted by $P_S y$, i.e., $P_S y = \arg\min_{x \in S} \|x - y\|$.



- Let us find an explicit expression of $P_S y$ in terms of a, b, and y.

Theorem: z is a solution of $\min_{x \in S} \|x - y\|$ if and only if

$$z \in S \text{ and } \langle z - y, z - x \rangle = 0 \quad \forall x \in S.$$

Proof: ① We first prove that: If $z \in S$ is a solution of $\min_{x \in S} \|x - y\|$, then $\langle z - y, z - x \rangle = 0 \quad \forall x \in S$.

Since z is a solution, $z \in S$, i.e., $\langle a, z\rangle = b$.

$\forall x \in S$ and $t \in \mathbb{R}$, it is easy to see that

$\langle a, (1+t)z - tx\rangle = (1+t)\langle a, z\rangle - t\langle a, x\rangle = b$.

Therefore, $(1+t)z - tx \in S$.

Since z is closest to y on S, we have

$$\|z-y\|^2 \le \|(1+t)z - tx - y\|^2$$
$$= \|(z-y) + t(z-x)\|^2$$
$$= \|z-y\|^2 + t^2\|z-x\|^2 + 2t\langle z-y, z-x\rangle.$$

i.e., $t\langle z-y, z-x\rangle \ge -\frac{t^2}{2}\|z-x\|^2$.

- If we choose $t > 0$,

$$\langle z-y, z-x\rangle \ge -\frac{t}{2}\|z-x\|^2$$

Letting $t \to 0_+$ gives $\langle z-y, z-x\rangle \ge 0$.

- If we choose $t < 0$,

$$\langle z-y, z-x\rangle \le -\frac{t}{2}\|z-x\|^2$$

Letting $t \to 0_-$ gives $\langle z-y, z-x\rangle \le 0$

Altogether, z satisfies $\langle z-y, z-x\rangle = 0 \quad \forall x \in S$.

② We then show that $z \in S$ satisfies $\langle z-y, z-x\rangle = 0$, then z is a solution of $\min_{x \in S} \|x-y\|$, by direct calculation.

Since $\langle z-y, z-x\rangle = 0 \quad \forall x \in S$,

$$\|x-y\|^2 = \|(z-x)-(z-y)\|^2$$
$$= \|z-x\|^2 + \|z-y\|^2 - 2\langle z-x, z-y\rangle$$
$$= \|z-y\|^2 + \|z-x\|^2 \ge \|z-y\|^2 \quad \forall x \in S.$$

This, together with $z \in S$, implies z minimizes $\|x-y\|^2$ in $x \in S$.

⊗

Theorem: The solution of $\min_{x\in S} \|x-y\|^2$ exists and unique, which is given by $y-\left(\frac{\langle a,y\rangle - b}{\|a\|^2}\right)a$

proof. denote $z = y-\left(\frac{\langle a,y\rangle - b}{\|a\|^2}\right)a$.

① $\langle a, z\rangle = \langle a,y\rangle - \left(\frac{\langle a,y\rangle - b}{\|a\|^2}\right)\langle a,a\rangle$

$= \langle a,y\rangle - (\langle a,y\rangle - b) = b$, so $z\in S$.

② $\forall\, x\in S$,

$\langle z-y, z-x\rangle = -\frac{\langle a,y\rangle - b}{\|a\|^2}\langle a, z-x\rangle$

$= -\frac{\langle a,y\rangle - b}{\|a\|^2}(\langle a,z\rangle - \langle a,x\rangle) = 0$

(because $\langle a,z\rangle = \langle a,x\rangle = b$).

By the previous theorem, z is a solution of $\min_{x\in S}\|x-y\|$.

It remains to show the uniqueness.

Suppose we have two solutions z_1 and z_2. Then,

z_1 is a solution, $\Rightarrow \langle z_1-y, z_1-z_2\rangle = 0$

z_2 is a solution, $\Rightarrow \langle z_2-y, z_2-z_1\rangle = 0$

Taking difference leads to $\langle z_1-z_2, z_1-z_2\rangle = 0$

$\Rightarrow \|z_1-z_2\|^2 = 0 \Rightarrow z_1 = z_2$ ⊠.

- In summary, the projection $P_S y$ of $y\in V$ onto the hyperplane

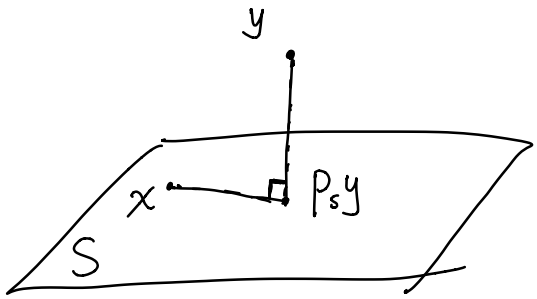
$$S = \{x\in V \mid \langle a,x\rangle = b\}$$

exists and is unique. Furthermore,

$$P_S y = y-\left(\frac{\langle a,y\rangle - b}{\|a\|^2}\right)a$$

and it satisfies

$$\langle P_S y - y, P_S y - x\rangle = 0.$$



Affine functions

A linear function plus a constant is called an affine function.
That is, a function $f: V \to \mathbb{R}$ is affine if

$$f(x) = g(x) + b,$$

where $g: V \to \mathbb{R}$ is linear and $b \in \mathbb{R}$ is a constant.

Properties:

- If $f: V \to \mathbb{R}$ is affine, then
  $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \quad \forall\, x, y \in V$ and $\alpha, \beta \in \mathbb{R}$ s.t. $\alpha + \beta = 1$.

  To see this, (g linear; $\alpha+\beta=1$)

  $$f(\alpha x + \beta y) = g(\alpha x + \beta y) + b = \alpha g(x) + \beta g(y) + (\alpha+\beta) b$$
  $$= \alpha(g(x)+b) + \beta(g(y)+b) = \alpha f(x) + \beta f(y).$$

- If $f: V \to \mathbb{R}$, where V is a Hilbert space, then
  f must be in the form of
  $$f(x) = \langle a, x \rangle + b, \quad \text{where } a \in V \text{ and } b \in \mathbb{R}.$$

§4.2 Case Studies: Regression and Classification.

§4.2.1 Linear Regression:

- Given a set of data

  $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$,

  where

  $x_i \in \mathbb{R}^n$ is an input feature vector
  $y_i \in \mathbb{R}$ is the corresponding response to x_i.
  , $i = 1, 2, \cdots, N$.

  Given a new input feature vector $x \in \mathbb{R}^n$, how to predict the corresponding response $y \in \mathbb{R}$?

  For example, $x_i \in \mathbb{R}^n$ represents n attributes of a house, and $y_i \in \mathbb{R}$ is the selling price.

  We want to predict the selling price of a house with feature $x \in \mathbb{R}^n$.

- Mathematically, we need to find a function $f: \mathbb{R}^n \to \mathbb{R}$ such that

  $$f(x_i) \simeq y_i \quad , \quad i = 1, 2, \cdots, N$$

  This is called regression. In this context,

  x_i, are called regressor / independent varibles.
  y_i are called dependent variables / outcome / label.

- The class of all functions $\mathbb{R}^n \to \mathbb{R}$ is too large, and the given data set $\{(x_i, y_i)\}_{i=1}^N$ is not enough to determine a function uniquely.

  So, we need to find a function class Φ where we search f.

  Intuitively, larger N, larger function class Φ.

- Linear model:

  We search f in the class of all affine functions,

  i.e., $f(x) = \langle a, x \rangle + b$ for some $a \in \mathbb{R}^n$, $b \in \mathbb{R}$.

Thus, we find $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$, s.t.

$$\langle a, x_i \rangle + b \approx y_i, \quad i = 1, 2, \cdots, N,$$

by minimizing the error of the linear equations.

While there are many possible definitions of error, it is popular to consider the square error as follows:

$$\left(\langle a, x_i \rangle + b - y_i\right)^2, \quad i = 1, 2, \cdots, N.$$

Therefore, we find $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ by solving

$$\min_{\substack{a \in \mathbb{R}^n, \\ b \in \mathbb{R}}} \sum_{i=1}^{N} \left(\langle a, x_i \rangle + b - y_i\right)^2$$

This problem is called the Least squares (LS) problem.

Write $X = \begin{bmatrix} x_1^T & 1 \\ x_2^T & 1 \\ \vdots & \vdots \\ x_N^T & 1 \end{bmatrix} \in \mathbb{R}^{N \times (n+1)}$, $\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}$

and $y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N$.

Then LS problem becomes

$$\boxed{\min_{\beta \in \mathbb{R}^{n+1}} \|X\beta - y\|_2^2 .}$$

Since we have N linear equations to fit and $n+1$ unknowns,

$N \geq n+1$ is required to have a unique solution.

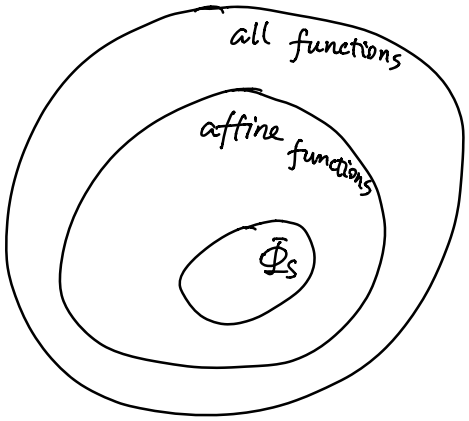
However, in practice, N can be much much smaller than n.

For example, x_i are images of size 10M pixels (i.e., $n = 10M$), and it is very difficult to have a data base of 10M images (i.e., $N << 10M$).

- Regularization.

  In many applications we don't have enough data (e.g., $N < n+1$), so that the class of affine functions is too large to search f.

  We search f is a subclass of all affine functions



Therefore, instead of searching $\beta \in \mathbb{R}^{n+1}$, we search $\beta \in S \subset \mathbb{R}^{n+1}$

$$\min_{\beta \in S} \|X\beta - y\|_2^2$$

In other words, we search f in

$$\Phi_S = \{ f : \mathbb{R}^n \to \mathbb{R} \mid f(x) = \langle a, x \rangle + b, \ \beta = \begin{bmatrix} a \\ b \end{bmatrix} \in S \}$$

By choosing different S, we obtain different approaches.

— Ridge regression:

We choose $S = \{ \beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1} \mid \|a\|_2 \le C \}$ for some $C > 0$.

The we solve $\min_{\beta \in S} \|X\beta - y\|_2^2$

$\Updownarrow$ ← by convex optimization theory

$$\min_{\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda \|a\|_2^2,$$

where $\lambda > 0$ depends on C and others.

Here $\|a\|_2^2$ is the regularization term.

and $\frac{1}{2}\|X\beta - y\|_2^2$ is the data-fitting term.

In other words, we find $\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}$ such that

the error of data fitting and the $\|a\|_2^2$

are minimized simultaneously.

Therefore, ridge regression gives $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$ such that

$X\beta \approx y$ and $\|a\|_2$ is small. (so that $\beta \in S$)

The parameter λ is tune s.t. $\|a\|_2 \le C$.

— LASSO regression.

We choose $S = \{\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1} \mid \|a\|_1 \le C\}$

The we solve $\min_{\beta \in S} \|X\beta - y\|_2^2$

$\Updownarrow$ ← by convex optimization theory

$$\min_{\beta = \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^{n+1}} \frac{1}{2}\|X\beta - y\|_2^2 + \lambda \|a\|_1$$

where $\lambda > 0$ depends on C and others.

Here the regularization term is $\|a\|_1$.

So, we find $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$ such that

the error of data fitting and the $\|a\|_1$

are minimized simultaneously.

Small $\|a\|_1$ tends to give a sparse vector $a \in \mathbb{R}^n$

(i.e., many entries of a are zeros)

Consequently, LASSO gives a solution $\beta = \begin{bmatrix} a \\ b \end{bmatrix}$ such that

$X\beta \approx y$ and a is sparse

Thus, given $x \in \mathbb{R}^n$, the prediction is given by

$$\langle x, a \rangle + b = \sum_{i=1}^{n} a_i x_i + b = \sum_{i \in S} a_i x_i + b$$

Let $I = \{i \mid a_i \neq 0\}$

Since I is a small set, the prediction depends only on

a small portion entries of x.

This is preferred because the prediction is interpretable.

§4.2.2. Kernel regression

Again, linear regression has its limitation.

We extend it to nonlinear regression by Kernel trick.

Feature map $\phi: \mathbb{R}^n \to H$ (H is some Hilbert space)

Then do regression in H.

However, since H is very large, the set of all linear functions is also too large. We need regularization.

We solve

$$\min_{a \in H} \frac{1}{2} \sum_{i=1}^{N} \left(\langle a, \phi(x_i) \rangle - y_i \right)^2 + \lambda \|a\|_H^2$$

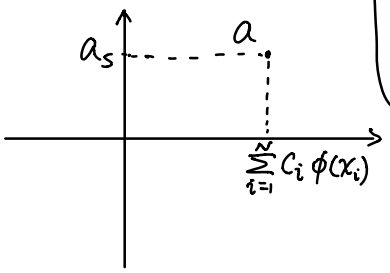
Representer Theorem:

The solution must be in the form of $a = \sum_{i=1}^{N} c_i \phi(x_i)$ for some $c = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$.

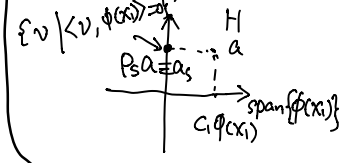
Proof. For any $a \in H$, we claim that a can be decomposed as

$$a = a_S + \sum_{i=1}^{N} c_i \phi(x_i)$$

where $c = \begin{bmatrix} c_1 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$ and $\langle a_S, \phi(x_i) \rangle = 0$ for $i = 1, 2, \cdots, N$.



Indeed, consider $N=1$ for simplicity.



Then, $S = \{v \mid \langle v, \phi(x_1) \rangle = 0\}$ is a hyperplane.

For any $a \in H$

$$a = P_S a + (a - P_S a),$$

where $P_S a$ is the projection of a onto S.

So, by the property of projection,

$$\langle P_S a - a, P_S a - 0 \rangle = 0 \quad (\text{because } 0 \in S)$$

$$\Rightarrow \langle P_S a - a, P_S a \rangle = 0$$

Also, by direct calculation,

$$a - P_S a = c_1 \phi(x_1)$$

So, $a = \underbrace{P_S a}_{a_S} + c_1 \phi(x_1)$

For general N, it can be done similarly.

Therefore,

$$\frac{1}{2}\sum_{i=1}^{N}\left(\langle a,\phi(x_i)\rangle - y_i\right)^2 + \lambda\|a\|_H^2$$

$$= \frac{1}{2}\sum_{i=1}^{N}\left(\langle \sum_{j=1}^{N} c_j\phi(x_j) + a_s,\ \phi(x_i)\rangle - y_i\right)^2 + \lambda\left\|\sum_{i=1}^{N} c_i\phi(x_i) + a_s\right\|_H^2$$

$$= \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} c_j\langle\phi(x_j),\phi(x_i)\rangle - y_i\right)^2 +$$

$$\lambda\left(\langle\sum_{i=1}^{N}c_i\phi(x_i),\sum_{j=1}^{N}c_j\phi(x_j)\rangle + 2\langle a_s,\sum_{i=1}^{N}c_i\phi(x_i)\rangle + \langle a_s,a_s\rangle\right)$$

$$= \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} c_j K(x_i,x_j) - y_i\right)^2 + \lambda\sum_{i=1}^{N}\sum_{j=1}^{N}c_ic_j K(x_i,x_j) + \lambda\|a_s\|_H^2$$

$$= \frac{1}{2}\|Kc-y\|_2^2 + \lambda c^TKc + \lambda\|a_s\|_H^2$$

where $K = \begin{bmatrix} K(x_1,x_1) & K(x_1,x_2) & \cdots & K(x_1,x_N) \\ \vdots & & & \\ K(x_N,x_1) & K(x_N,x_2) & \cdots & K(x_N,x_N) \end{bmatrix} \in \mathbb{R}^{N\times N}$ $\quad c = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{bmatrix} \in \mathbb{R}^N$.

Let $F_1(c) = \frac{1}{2}\|Kc-y\|_2^2 + \lambda c^TKc$ —— depends on $c\in\mathbb{R}^N$ only.

$F_2(a_s) = \lambda\|a_s\|_H^2$ —— depends on $a_s \in H$ only.

Then, the minimization is the same as

$$\min_{\substack{c\in\mathbb{R}^N \\ a_s\in H \\ \langle a_s,\phi(x_i)\rangle=0,\ i=1,\cdots,N}} F_1(c) + F_2(a_s)$$

$$\Updownarrow \quad a = a_s + \sum_{i=1}^{N} c_i\phi(x_i)$$

$$\min_{c\in\mathbb{R}^N} F_1(c) \quad \text{and} \quad \min_{\substack{a_s\in H \\ \langle a_s,\phi(x_i)\rangle=0 \\ i=1,\cdots,N}} F_2(a_s)$$

Obviously, because $\|a_s\|_H^2 \geq 0$, $\min_{\substack{a_s\in H \\ \langle a_s,\phi(x_i)\rangle=0}} F(a_s)$ is solved by $a_s = 0$.

Thus, the solution of the original minimization is

$$a = \sum_{i=1}^{N} c_i\phi(x_i)$$

where $c\in\mathbb{R}^N$ is a solution of $\min_{c\in\mathbb{R}^N} F_1(c)$. ⊠

From the above proof, we see that

$$\min_{a \in H} \frac{1}{2} \sum_{i=1}^{N} (\langle a, \phi(x_i) \rangle - y_i)^2 + \lambda \|a\|_H^2$$

$$\Updownarrow$$

$$\boxed{\min_{C \in \mathbb{R}^N} \frac{1}{2} \|KC - y\|_2^2 + \lambda C^T K C}$$

Let the solution be $C \in \mathbb{R}^N$.

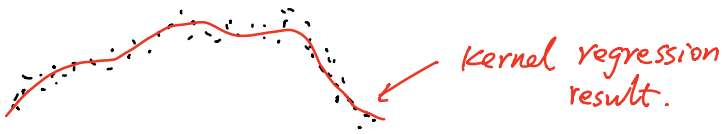
Then the predicted output y for input $x \in \mathbb{R}^n$ is

$$y = \langle a, \phi(x) \rangle = \langle \sum_{i=1}^{N} c_i \phi(x_i), \phi(x) \rangle$$

$$= \sum_{i=1}^{N} c_i K(x_i, x) .$$

All the computation involves only the kernel function $K(\cdot, \cdot)$,

No explicit feature map $\phi(\cdot)$ is needed.



§4.2.3 Classification

- Classification: Giving training data

$$(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N), \quad x_i \in \mathbb{R}^n, \quad y_i \in \{-1, +1\}, \quad i = 1, \cdots, N,$$

find a classifier (a function) f such that

$$y_i = \begin{cases} 1, & \text{if } f(x_i) \geq 1 \\ -1, & \text{if } f(x_i) \leq -1 \end{cases}.$$

We use hyperplanes to separate the points

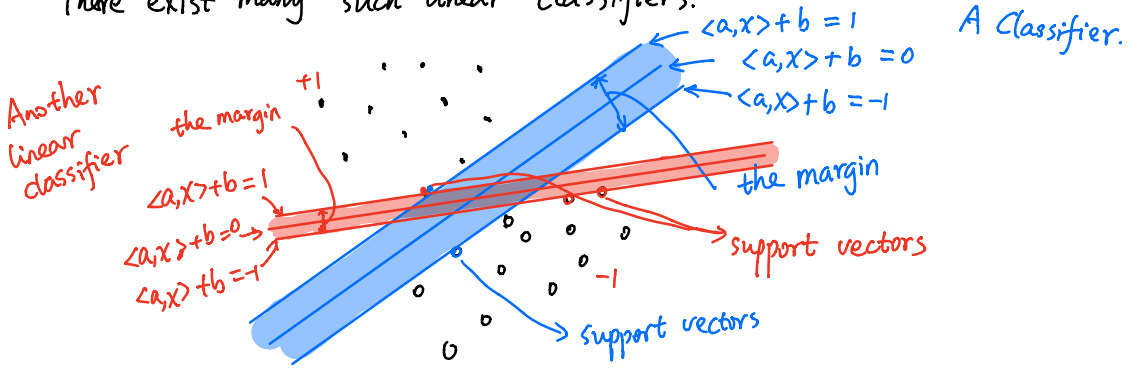
$$f(x) = \langle a, x \rangle + b, \quad \text{where } a \in \mathbb{R}^n, \quad b \in \mathbb{R}.$$

The weights $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$ are noralized such that

$$\langle a, x_i \rangle + b_i \begin{cases} \geq 1 & \text{if } y_i = +1 \\ \leq -1 & \text{if } y_i = -1 \end{cases}$$

- Support Vector Machine (SVM)

There exist many such linear classifiers.



Which one is the better?

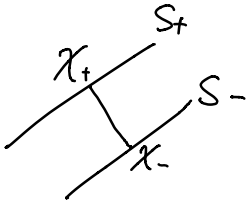
From the figure, the blue is better, because it has a larger margin, and hence a larger buffer zone of mis-classification.

Therefore, we want to maximize the margin among all candidates.

Let us calculate the margin in terms of a and b.

The margin is the distance between the two hyperplanes

$$S_+ = \{x \mid \langle a, x \rangle + b = 1\} \text{ and } S_- = \{x \mid \langle a, x \rangle + b = -1\}$$



Let $x_+ \in S_+$ and $x_- \in S_-$ such that

$$\|x_+ - x_-\|_2 = \text{dist}(S_+, S_-).$$

Since x_+ is a projection of x_- onto $S_+ = \{x \mid \langle a, x\rangle = 1-b\}$

$$\begin{aligned} x_+ &= x_- - \frac{\langle a, x_-\rangle + b - 1}{\|a\|_2^2} a \\ &= x_- - \frac{-1-b+b-1}{\|a\|^2} a \qquad (\text{since } x_- \in S_-) \\ &= x_- + \frac{2}{\|a\|_2^2} a \end{aligned}$$

Thus, $\|x_+ - x_-\|_2 = \left\|\frac{2}{\|a\|_2^2} a\right\|_2 = \frac{2}{\|a\|_2}$

i.e., the margin $= \frac{2}{\|a\|_2}$

Support Vector Machine (SVM)

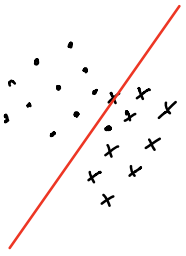
$$\begin{aligned} \max_{\substack{a\in\mathbb{R}^n \\ b\in\mathbb{R}}} \quad & \frac{2}{\|a\|_2} \\ \text{s.t.} \quad & \langle a, x_i\rangle + b \geq 1 \quad \text{if } y_i = 1 \\ & \langle a, x_i\rangle + b \leq -1 \quad \text{if } y_i = -1 \end{aligned},$$

which is equivalent to

$$\begin{aligned} \min_{a\in\mathbb{R}^n,\, b\in\mathbb{R}} \quad & \frac{1}{2}\|a\|_2^2 \\ \text{s.t} \quad & y_i(\langle a, x_i\rangle + b) \geq 1 \end{aligned} \qquad \text{(SVM-1)}$$

- The above SVM is NOT robust to noise.

For example shown on the right,
even we have only two noisy points, there is no solution to (SVM-1).

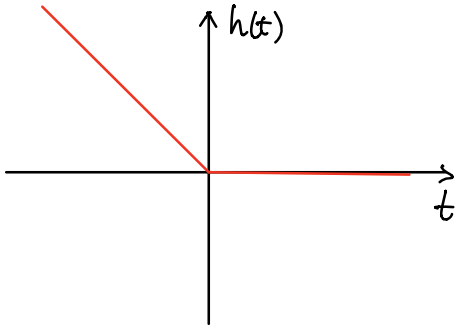


We consider a "soft" version of (SVM-1) as follows.

We minimize the error to the separation

$$\sum_{i=1}^{N} h\left(\left(y_i \langle a, x_i\rangle + b\right) - 1\right),$$

where the function $h(t) = \begin{cases} 0 & \text{if } t \geq 0 \\ |t| & \text{if } t < 0 \end{cases}$



In other words, if $y_i\langle a, x_i\rangle + b \geq 1$, the error is 0.

if $y_i\langle a, x_i\rangle + b < 1$, the error is the absolute value of $y_i\langle a, x_i\rangle + b_i - 1$.

Also, the margin $\frac{1}{2}\|a\|_2^2$ can be viewed as a regularization.

Altogether, we solve

$$\min_{\substack{a\in\mathbb{R}^n \\ b\in\mathbb{R}}} \sum_{i=1}^{N} h\left(y_i\langle a, x_i\rangle + b_i - 1\right) + \frac{1}{2}\|a\|_2^2 \qquad (SVM\text{-}2)$$

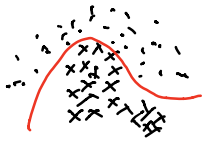
We call this SVM with a soft margin.

— h can be approximated by some smooth function. If h is chosen the so-called logistic function, we call it logistic regression.

- Kernel SVM

The linear SVMs don't work for curved data.

The kernel method can be used.



Let $\phi : \mathbb{R}^n \to H$ be a feature map.

So, we use linear functions on H to classify the points

By Riesz representation theorem, any linear function in the form of $\langle a, x\rangle$ for some $a \in H$.

Therefore, (SVM-2) becomes

$$\min_{a\in H} h\left(y_i\langle a, \phi(x_i)\rangle - 1\right) + \frac{1}{2}\|a\|_H^2 \qquad (K\text{-}SVM)$$

Aagain, one can prove the following representer theorem.

Theorem: Any solution of (K-SVM) is in the form of

$$a = \sum_{i=1}^{N} c_i \phi(x_i)$$

proof. Write $a = \sum_{i=1}^{N} c_i \phi(x_i) + a_S$ for some $a_S \in H$ and $\langle a_S, \phi(x_i) \rangle = 0 \ \forall i$.

The rest is the same as the linear regression case. ⊠.

Thus, (K-SVM) becomes

$$\min_{c \in \mathbb{R}^N} h\left(y_i \left(\sum_{j=1}^{N} K(x_i, x_j) c_j \right) - 1 \right) + \frac{1}{2} c^T K c,$$

where $K = [K(x_i, x_j)]_{i=1, j=1}^{N, N} \in \mathbb{R}^{N \times N}$.

The prediction of the input x is given by

$$\operatorname{sgn}\left(\sum_{j=1}^{N} K(x, x_j) c_j \right)$$

Again, only $K(\cdot, \cdot)$ is needed in the kernel SVM, and no explicit feature maps $\phi(\cdot)$ is required.

§4.3. Linear approximation and Differentiation.

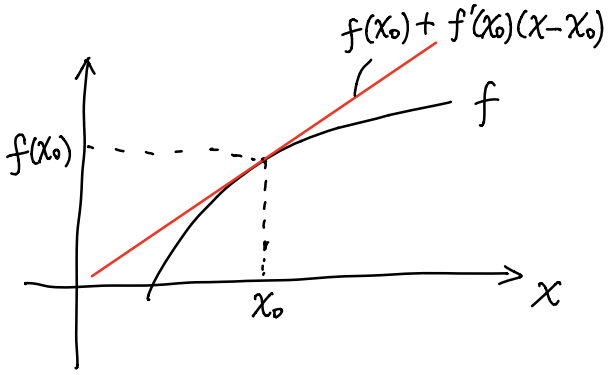
Recall that for a function $f: \mathbb{R} \to \mathbb{R}$, the derivative at x_0 is

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0},$$

which is the same as

$$\lim_{x \to x_0} \left| \frac{f(x) - f(x_0) - f'(x_0)(x - x_0)}{x - x_0} \right| = 0$$

Notice that $f(x_0) + f'(x_0)(x - x_0)$ is an affine function in $\mathbb{R}$ that passes through $(x_0, f(x_0))$.



In other words, in differentiation at x_0,

(1) f is approximated by an affine function passes thru $(x_0, f(x_0))$.

(2) the error of the approximation is $o(|x - x_0|)$.

(o: little o, i.e., $\lim_{x \to x_0} \frac{|\text{error}|}{|x - x_0|} \to 0$)

This can be used to define differentiation of functions on Hilbert spaces.
(We use Hilbert space for simplicity. It can be easily adapted to Banach spaces.)

Let $f: V \to \mathbb{R}$ with V a Hilbert space.

Consider the differentiation of f at $x^{(0)} \in V$.

(1) By Riesz representation theorem, any affine function is in the form

of $\langle v, x\rangle + a$ for some $v \in V$ and $a \in \mathbb{R}$. Since it passes thru $(x^{(0)}, f(x^{(0)}))$, $\langle v, x^{(0)}\rangle + a = f(x^{(0)})$ Therefore, the affine function is in the form of

$$\langle v, x\rangle + a = \langle v, x - x^{(0)}\rangle + (\langle v, x^{(0)}\rangle + a)$$
$$= f(x^{(0)}) + \langle v, x - x^{(0)}\rangle.$$

(2). The approxition error is

$$\text{error} = \left| f(x) - f(x^{(0)}) - \langle v, x - x^{(0)}\rangle \right|,$$

The error should be in the order of $o(\|x - x^{(0)}\|)$, i.e.,

$$\frac{\text{error}}{\|x - x^{(0)}\|} \to 0 \quad \text{as} \quad x \to x^{(0)} \ (\text{i.e., } \|x - x^{(0)}\| \to 0)$$

Definition: Let V be a Hilbert space. Let $f: V \to \mathbb{R}$. Then f is said Frechet differentiable if there exists a $v \in V$ such that

$$\lim_{x \to x^{(0)}} \frac{\left| f(x) - f(x^{(0)}) - \langle v, x - x^{(0)}\rangle \right|}{\|x - x^{(0)}\|} = 0. \quad \left(\begin{array}{c}\text{Note that } x \to x^{(0)} \text{ is the} \\ \text{same as } \|x - x^{(0)}\| \to 0\end{array}\right)$$

If f is differentiable at $x^{(0)}$, v is called the gradient of f at $x^{(0)}$, denoted by $\nabla f(x^{(0)})$.

Example 1: $f(x) = \|x\|^2$, where $\|x\|$ is the norm on V.

At any $x^{(0)} \in V$,

$$\|x\|^2 = \|(x - x^{(0)}) + x^{(0)}\|^2 = \langle (x - x^{(0)}) + x^{(0)}, (x - x^{(0)}) + x^{(0)}\rangle$$
$$= \|x - x^{(0)}\|^2 + \|x^{(0)}\|^2 + 2\langle x^{(0)}, x - x^{(0)}\rangle$$

Therefore, $\|x\|^2 - \underbrace{\left(\|x^{(0)}\|^2 + 2\langle x^{(0)}, x - x^{(0)}\rangle\right)}_{\downarrow \text{ affine approximation}} = \|x - x^{(0)}\|^2$

So $$\lim_{x \to x^{(0)}} \frac{\left| \|x\|^2 - \|x^{(0)}\|^2 - \langle 2x^{(0)}, x - x^{(0)}\rangle \right|}{\|x - x^{(0)}\|} = \lim_{x \to x_0} \frac{\|x - x^{(0)}\|^2}{\|x - x^{(0)}\|} = 0.$$

Thus, $\nabla f(x^{(0)}) = 2x^{(0)}$.

Example 2: $f(x) = \langle a, x \rangle$ for some $a \in V$.

At any $x^{(0)} \in V$,

$$\langle a, x \rangle = \langle a, x^{(0)} \rangle + \langle a, x - x^{(0)} \rangle$$

Therefore, $\lim_{x \to x^{(0)}} \frac{|\langle a, x \rangle - \langle a, x^{(0)} \rangle - \langle a, x - x^{(0)} \rangle|}{\|x - x^{(0)}\|} = \lim_{x_0 \to x} \frac{0}{\|x - x^{(0)}\|} = 0.$

Thus, $\nabla f(x^{(0)}) = a$.

Example 3: $f(x) = \|x - a\|^2$ for some $a \in V$.

At any $x^{(0)} \in V$,

$$f(x) = \|x - a\|^2 = \|(x^{(0)} - a) + (x - x^{(0)})\|^2$$
$$= \|x^{(0)} - a\|^2 + \|x - x^{(0)}\|^2 + 2\langle x^{(0)} - a, x - x^{(0)} \rangle$$
$$= f(x^{(0)}) + \langle 2(x^{(0)} - a), x - x^{(0)} \rangle + \|x - x^{(0)}\|^2$$

So, $\lim_{x \to x^{(0)}} \frac{|f(x) - f(x^{(0)}) - \langle 2(x^{(0)} - a), x - x^{(0)} \rangle|}{\|x - x^{(0)}\|} = \lim_{x \to x^{(0)}} \|x - x^{(0)}\| = 0.$

Therefore, $\nabla f(x^{(0)}) = 2(x^{(0)} - a)$

Properties:

① Frechet differentiation is linear (in f), i.e.,

$$\nabla(\alpha f + \beta g)(x) = \alpha \nabla f(x) + \beta \nabla g(x).$$

② Chain rule: Let $f: V \to \mathbb{R}$ and $g: \mathbb{R} \to \mathbb{R}$. Then $g \circ f: V \to \mathbb{R}$ and

$$\nabla(g \circ f)(x) = g'(f(x)) \cdot \nabla f(x)$$

if f and g are differentiable at x and $f(x)$ respectively.

Example 4: $f(x) = \|x\| \quad \forall\, x \in V$.

This is a composition of $f_1(x) = \|x\|^2$ from $V \to \mathbb{R}$

and $f_2(t) = \sqrt{t}$ from $\mathbb{R} \to \mathbb{R}$.

When $\|x\| \neq 0$, both f_1 and f_2 are differentiable.

Also, $\nabla f_1(x) = 2x$, $f_2'(t) = \frac{1}{2\sqrt{t}}$ if $t \neq 0$.

So, $\nabla f(x) = \nabla (f_2 \circ f_1)(x) = f_2'(f_1(x)) \cdot \nabla f_1(x)$

$$= \frac{1}{2\sqrt{\|x\|^2}} \cdot 2x = \frac{x}{\|x\|}.$$

When $\|x\| = 0$, (i.e., $x = 0$), $f_2(t)$ is NOT differentiable at $f_1(x) = 0$.

It can be shown that $f(x) = \|x\|$ is NOT differentiable at $x = 0$.

③ For functions on $\mathbb{R}^n$: $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \frac{\partial f}{\partial x_2}(x) \\ \vdots \\ \frac{\partial f}{\partial x_n}(x) \end{pmatrix}, \quad \text{where} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

i.e., Frechet differentiation is the same as the standard differentiation in multivariate calculus.

Taylor's expansion

From the definition, we see that

$$f(x) \approx f(x^{(0)}) + \langle \nabla f(x^{(0)}), x - x^{(0)} \rangle$$

Or, more precisely,

$$f(x) = f(x^{(0)}) + \langle \nabla f(x^{(0)}), x - x^{(0)} \rangle + o(\|x - x^{(0)}\|)$$

This is a generalization of Taylor's expansion.

In paticular, if $f : \mathbb{R}^n \to \mathbb{R}$

$$f(x) \approx f(x^{(0)}) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^{(0)}) \, (x_i - x_i^{(0)}) \, .$$

Differentiation on normed vector spaces

Let V be a normed vector space.

Let $f: V \to \mathbb{R}$ be a function. Let $x^{(0)} \in V$.

To define differentiation, we still use an affine function approximation, and the affine function passes thru $(x^{(0)}, f(x^{(0)}))$.

Thus, we use $f(x^{(0)}) + L(x - x^{(0)})$, where $L: V \to \mathbb{R}$ is a linear function, to approximate $f(x)$.

However, because there is no Riesz representation on normed spaces, we keep the linear function L in the definition of differentiation.

Definition: f is differentiable at $x^{(0)} \in V$, if:

$\exists$ a linear function $L: V \to \mathbb{R}$ such that

$$\lim_{\|x - x^{(0)}\| \to 0} \frac{\left| f(x) - \left(f(x^{(0)}) + L(x - x^{(0)})\right) \right|}{\|x - x^{(0)}\|} = 0 .$$

The linear function L is called the differentiation of f at $x^{(0)}$.