Artificial Intelligence and Machine Learning

Mathematical Foundations

UMC 203

Author

Nitin Vetcha

Instructor

Professor Chiranjib Bhattacharyya

Department of Computer Science and Automation

Contents

I

Pre Midsem

1 Introduction to Classification, Bayes Classifier, Multivariate Gaussian and Optimality

In Binary Classification, we are given be a pair of random variables (X, Y) where X is called Instance Space and Y is is known as Label Space, taking their respective values from $\mathbb{R}^d$ and $\{-1, 1\}$ i.e., $X \subseteq \mathbb{R}^d$ and $Y = \{-1, 1\}$. Any function $h : X \mapsto \{-1, 1\}$ is known as a Classifier. We normally refer to the probabilities

- $P_1 = P(Y = 1)$ and $P_2 = P(Y = -1)$ as Priori Probabilities (or simply Prior)
- $P(X = x \mid Y = 1)$ and $P(X = x \mid Y = -1)$ as Class Conditional Distributions. These can either be densities (probability density function) or probability mass function
- $P(Y = 1 \mid X = x) = \eta(x)$ and $P(Y = -1 \mid X = x) = 1 - \eta(x)$, since probabilities must always add up to 1, as Posteriori Probabilities (or simply Posterior)

Posteriori probabilities can be rewritten in terms of the Priori probabilities and class conditional distribution using the Bayes Formula as follows,

$$P(Y = 1 \mid X = x) = \frac{P(Y = 1)\, P(X = x \mid Y = 1)}{P(X = x)}$$

§1.1 Bayes Classifier

Lets consider the classifier $h(x)$ given as follows,

$$h(x) = \begin{cases} 1 & P(Y = 1 \mid X = x) > P(Y = -1 \mid X = x) \\ -1 & P(Y = 1 \mid X = x) \leq P(Y = -1 \mid X = x) \end{cases}$$

$$\implies h(x) = \begin{cases} 1 & \eta(x) > 1 - \eta(x) \\ -1 & \eta(x) \leq 1 - \eta(x) \end{cases}$$

If we define $sign(z)$ as follows,

$$sign(z) = \begin{cases} 1 & z > 0 \\ -1 & z \leq 0 \end{cases}$$

then

$$h^*(x) = sign(2\eta(x) - 1)$$

is known as the Bayes Classifier.

§1.2 Specific Case : Multivariate Gaussian

We shall consider the specific case,

$$P(X = x \mid Y = 1) = N(x \mid \mu_1, \Sigma_1)$$
$$P(X = x \mid Y = -1) = N(x \mid \mu_2, \Sigma_2)$$

where $N(x \mid \mu, \Sigma)$ is the Multivariate Normal Density in d dimensions whose P.D.F is given by

$$N(x \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi}\,|\Sigma|^{d/2}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

Here $x \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d\times d}$. Σ also known as covariance matrix is always symmetric and positive semi-definite. ($|\Sigma|$ and Σ^{-1} denote respectively the determinant and inverse of Σ).

Considering $\Sigma_1 = \Sigma_2 = \Sigma$, we have

$$\eta(x) = \frac{P_1\, N(x \mid \mu_1, \Sigma)}{P(X = x)}$$

In this case, the Bayes Classifier can be given as

$$h^*(x) = \begin{cases} 1 & \log\left(\dfrac{\eta(x)}{1-\eta(x)}\right) > 0 \\ -1 & \text{otherwise} \end{cases}$$
$$h^*(x) = sign\left[\log\left(\frac{\eta(x)}{1-\eta(x)}\right)\right]$$

Lets simplify the term inside the bracket.

$$\log\left(\frac{\eta(x)}{1-\eta(x)}\right) = \log\left(\frac{P_1}{P_2}\frac{N(x \mid \mu_1, \Sigma)}{N(x \mid \mu_2, \Sigma)}\right)$$
$$= \log\left(\frac{P_1}{P_2}\right) + \log\left(\frac{N(x \mid \mu_1, \Sigma)}{N(x \mid \mu_2, \Sigma)}\right)$$

Now, lets expand $\log\left(\dfrac{N(x \mid \mu_1, \Sigma)}{N(x \mid \mu_2, \Sigma)}\right)$

$$= -\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)$$
$$= \frac{1}{2}\left[(\cancel{-x^T\Sigma^{-1}x} + x^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}x - \mu_1^T\Sigma^{-1}\mu_1) + (\cancel{x^T\Sigma^{-1}x} - x^T\Sigma^{-1}\mu_2 - \mu_2^T\Sigma^{-1}x + \mu_2^T\Sigma^{-1}\mu_2)\right]$$
$$= \frac{1}{2}\left[(\mu_1^T - \mu_2^T)\Sigma^{-1}x + x^T\Sigma^{-1}(\mu_1 - \mu_2) + \mu_2^T\Sigma^{-1}\mu_2 - \mu_1^T\Sigma^{-1}\mu_1\right]$$
$$= (\mu_1^T - \mu_2^T)\Sigma^{-1}x - \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2\right)$$

The last equality is due to the fact that for a scalar p, we have $p = p^T$ and $(\mu_1^T - \mu_2^T)\Sigma^{-1}x$ is a scalar (since its dimension is $1\times 1 = (1\times d)\cdot(d\times d)\cdot(d\times 1)$), thus $(\mu_1^T - \mu_2^T)\Sigma^{-1}x = ((\mu_1^T - \mu_2^T)\Sigma^{-1}x)^T = x^T(\Sigma^{-1})^T(\mu_1 - \mu_2) = x^T\Sigma^{-1}(\mu_1 - \mu_2)$. We thus have that,

$$\log\left(\frac{\eta(x)}{1-\eta(x)}\right) = (\mu_1^T - \mu_2^T)\Sigma^{-1}x - \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2\right) + \log\left(\frac{P_1}{P_2}\right)$$
$$= w^T x + b$$

Note that,

$$w = \Sigma^{-1}(\mu_1 - \mu_2)$$
$$b = \log\left(\frac{P_1}{P_2}\right) - \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2\right)$$

Thus, $h^*(x) = sign(w^T x + b)$ which shows that the Bayes classifier is a linear function. Similarly, examine the cases where $\Sigma_1 = \sigma_1^2 I$ and $\Sigma_2 = \sigma_2^2 I$ as well as where $\Sigma_1 \neq \Sigma_2$. The equation $w^T x = -b$ defines a hyperplane while the regions denoted by $w^T x > b$ and $w^T x < b$ are referred to as halt spaces.

§1.3 Optimality

We shall now address the question : How good is the Bayes Classifier ? Before proceeding, note that if for an event A, we denote the indicator variable by 1_A, then we have that $E(1_A) = P(A)$. Using this, we have that for a classifier h

$$P(h(X) \neq Y) = E_{XY}(1_{h(X)\neq Y})$$
$$= E_X E_{Y/X}(1_{h(X)\neq Y})$$

Since

$$E_{Y/X}(1_{h(X)\neq Y}) = \eta(x)(1_{h(x)\neq 1}) + (1-\eta(x))(1_{h(x)\neq -1})$$

we have that

$$P(h(X) \neq Y) = E_X(\eta(x)(1_{h(x)\neq 1}) + (1-\eta(x))(1_{h(x)\neq -1}))$$

Thus,

$$P(h(X) \neq Y) - P(h^*(X) \neq Y) = E_X(E_{Y/X}(1_{h(X)\neq Y} - 1_{h^*(X)\neq Y}))$$

Noting that, $h(X) \neq h^*(X)$ implies that if $h(X) \neq Y$ then $h^*(X) = Y$, we have

$$\begin{aligned}
E_{Y/X}(1_{h(X)\neq Y} - 1_{h^*(X)\neq Y}) &= E_{Y/X}(1_{h^*(X)=Y} - 1_{h^*(X)\neq Y}) \\
&= E_{Y/X}(2\cdot 1_{h^*(X)=Y} - 1) \\
&= \eta(x)(2\cdot 1_{h^*(x)=1} - 1) + (1-\eta(x))(2\cdot 1_{h^*(x)=-1} - 1) \\
&= \eta(x)(2\cdot 1_{h^*(x)=1} - 1) + (1-\eta(x))(2\cdot(1 - 1_{h^*(x)=1}) - 1) \\
&= (2\eta(x) - 1)(2\cdot 1_{h^*(X)=1} - 1) \\
&= |2\eta(x) - 1|
\end{aligned}$$

The last equality can be shown to be true if we observe the below table.

$\eta(x)$	$2\eta(x)-1$	$2\cdot 1_{h^*(X)=1} - 1$
$> (1/2)$	> 0	> 0
$\leq (1/2)$	≤ 0	≤ 0

Thus,

$$P(h(X) \neq Y) - P(h^*(X) \neq Y) = E_X(|2\eta(x) - 1) \geq 0$$

which shows that Bayes classifier is the best classifier.

2 Bayes Error Rate and Discriminant Functions

We had previously seen that,

$$\begin{aligned} E_{Y/X}(1_{h(X)\neq Y}) &= \eta(x)(1_{h(x)\neq 1}) + (1-\eta(x))(1_{h(x)\neq -1}) \\ &= \begin{cases} 1-\eta(x) & h(x)=1 \\ \eta(x) & h(x)=-1 \end{cases} \end{aligned}$$

We say that there is a mistake if $Y = -1$ but $h(X) = 1$ and if $Y = 1$ but $h(X) = -1$. If $\eta(x) > (1-\eta(x))$, then predict $h(x) = 1$. The minimum over h (all possible classifiers) is attained by choosing $h(x)$ such that

$$E_{Y\,|\,X=x} = \min(\eta(x), 1-\eta(x))$$

$$h^*(X) = \begin{cases} 1 & \eta(x) > 1-\eta(x) \\ -1 & \eta(x) \leq 1-\eta(x) \end{cases}$$

§2.1 Bayes Decision Theory

The setting is same as before i.e, we have an instance space $x \subseteq \mathbb{R}^d$ with label $y \in \{-1, 1\}$ and we predict $\hat{y} \in \{-1, 1\}$. Given a loss function $l(\hat{y}, y) : Y \times Y \mapsto \mathbb{R}_+$, we define expected loss as

$$R(h) = E_{X,Y}(l(h(X), Y))$$

We are interested in finding

$$\min_h R(h)$$

Suppose

$$\begin{aligned} R(\tilde{h}) &= \min_h R(h) \\ \implies \tilde{h}(x) &= \min_h E_{Y\,|\,X=x}(l(h(X), Y)) \end{aligned}$$

Then, $R(\tilde{h})$ is called the Bayes Error Rate. For a given instance, it chooses the label which yields minimum loss. Lets consider the 2 class problem where $Y = \{-1, 1\}$.

$$\begin{aligned} E_{Y\,|\,X=x}(l(1, Y)) &= l(1,1)P(Y=1 \mid X=x) + l(1,-1)P(Y=-1 \mid X=x) \\ &= l(1,1)\eta(x) + l(1,-1)(1-\eta(x)) \\ E_{Y\,|\,X=x}(l(-1, Y)) &= l(-1,1)\eta(x) + l(-1,-1)(1-\eta(x)) \end{aligned}$$

Now, choose h such that $E_{Y\,|\,X=x}\,(l(h(X), Y))$ should be minimum.

- Choose class $\tilde{h}(x) = 1$ if $E_{Y\,|\,X}\,(l(1, Y)) < E_{Y\,|\,X}(l(-1, Y))$
- Choose class $\tilde{h}(x) = -1$ if $E_{Y\,|\,X}\,(l(-1, Y)) < E_{Y\,|\,X}(l(1, Y))$

$$\begin{aligned}
\tilde{h}(x) = 1 &\implies l(1,1)\eta(x) + l(1,-1)(1-\eta(x)) < l(-1,1)\eta(x) + l(-1,-1)(1-\eta(x)) \\
&\implies (l(1,1) - l(-1,1))\,\eta(x) < (l(-1,-1) - l(1,-1))\,(1-\eta(x)) \\
\tilde{h}(x) = -1 &\implies l(-1,1)\eta(x) + l(-1,-1)(1-\eta(x)) < l(1,1)\eta(x) + l(1,-1)(1-\eta(x)) \\
&\implies (l(-1,-1) - l(1,-1))\,(1-\eta(x)) < (l(1,1) - l(-1,1))\,\eta(x)
\end{aligned}$$

Properties of l

- It should be non-negative
- It should penalize mistakes more

If we choose

$$\begin{aligned}
l(-1,1) = l(1,-1) = 1 \\
l(1,1) = l(-1,-1) = 0
\end{aligned}$$

then we have that,

$$\begin{aligned}
\tilde{h}(x) = 1 \quad &\text{if } \eta(x) > 1 - \eta(x) \\
\tilde{h}(x) = -1 \quad &\text{if } 1 - \eta(x) > \eta(x)
\end{aligned}$$

Thus,

$$\tilde{h}(x) = h^*(x)$$

§2.2 Multi-Category Classification

We have the following notations in this case

- X is the Instance Space and Y the label space
- Classifier $h : X \mapsto \{1, \ldots, M\}$

For $i \in \{1, \ldots, M\}$

- Prior $P_i = P(Y = i)$
- Class conditional distribution $P(X = x \mid Y = i)$ which are either densities or p.m.f
- Posterior $\eta_i(x) = P(Y = i \mid X = x)$
- $P(Y = i \mid X = x) = \dfrac{P_i P(X = x \mid Y = i)}{\sum_k (P_k P(X = x \mid Y = k))}$

Bayes classifier is given as,

$$h^*(x) = i \text{ if } P(Y = i \mid X = x) > P(Y = j \mid X = x) \text{ for all } j \neq i$$

Bayes Error Rate is given by

$$R(\tilde{h}) = \min_{h \in H} R(h)$$

where

$$\begin{aligned}
R(h) &= P(h(X) \neq Y) \\
&= E(l(h(X), Y)) \\
\tilde{h} &= \arg\min_h E_{Y/X}\left(l(h(X), Y)\right)
\end{aligned}$$

Here

$$E_{Y|X=x}\left(l(h(X), Y)\right) = \min_{i\in\{1,\ldots,M\}} r_i(x)$$

where

$$r_i(x) = E_{Y|X=x}(l(i, Y))$$

For every x, choose $\tilde{h}(x) = i$ if $r_i(x) < r_j(x)$ for all $j \neq 1$. Now, choose

$$\begin{aligned} l(i, i) &= 0 \quad i \in \{1, \ldots, M\} \\ l(i, j) &= 1 \qquad\qquad i \neq j \end{aligned}$$

Thus,

$$\begin{aligned} r_i(x) &= \sum_{k=1}^{M} l(i, k)\, \eta_k(x) \\ &= l(i, i)\, \eta_i(x) + \sum_{k\neq i}^{M} l(i, k)\, \eta_k(x) \\ &= \sum_{k\neq i}^{M} l(i, k)\, \eta_k(x) \\ &= 1 - \eta_i(x) \end{aligned}$$

Hence, for all $j \neq i$, we have that

$$\begin{aligned} r_i(x) < r_j(x) &\implies 1 - \eta_i(x) < 1 - \eta_j(x) \\ &\implies \eta_j(x) < \eta_i(x) \end{aligned}$$

Thus, we are choosing $\tilde{h}(x) = i$ for all $j \neq i$ when $\eta_j(x) < \eta_i(x)$. Hence, even in this case, we have that

$$\tilde{h}(x) = h^*(x)$$

§2.3 Discriminant Functions

For the M category problem, we define discriminant functions $g_i : , X \mapsto \mathbb{R}$ for $i \in \{1, \ldots, M\}$. We now define a classifier h, such that if $g_i(x) > g_j(x)$ for all $j \neq i$, then $h(x) = i$. If $f : [0, 1] \mapsto \mathbb{R}$ is any monotonically increasing function, then $g_i(x) = f(\eta_i(x))$ can be a discriminant function.

Lets consider the case where

$$P(X = x \mid Y = i) = N(x \mid \mu_i, \Sigma_i)$$

then

$$\eta_i(x) = \frac{P_i N(x \mid \mu_i, \Sigma_i)}{P(x)}$$

If we now assume that $\Sigma_i = \Sigma$ for $i \in \{1, \ldots, M\}$, then we have that

$$\eta_i(x) = \frac{P_i N(x \mid \mu_i, \Sigma)}{P(x)}$$

Taking $f(z) = \log(z)$ where $0 \leq z \leq 1$, we have that

$$\begin{aligned}\log(\eta_i(x)) &= \log(P_i) + \log(N(x \mid \mu_i, \Sigma)) - \log(P(x)) \\ &= \log(P_i) + \log\left(\frac{1}{(\sqrt{2\pi})^d |\Sigma|^{1/2}}\right) - \frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i) - \log(P(x))\end{aligned}$$

Excluding constant terms, we have

$$g_i(x) = \log(P_i) - \frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i)$$

Then Bayes Classifier is given as $h^*(x) = i$ if $g_i(x) > g_j(x)$ for all $j \neq i$ and $i, j \in \{1, \ldots, M\}$.

If $M = 2$, then

$$h^*(x) = \begin{cases} 1 & g_1(x) - g_2(x) > 0 \\ 2 & g_1(x) - g_2(x) < 0 \end{cases}$$

Alternatively, we can also have $h^*(x) = sign(g(x))$ where

$$\begin{aligned} g_x &= g_1(x) - g_2(x) \\ &= \log(P_1) - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) - \log(P_2) - \frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2) \\ &= w^T x + b \end{aligned}$$

Now, if $P_1 = P_2 = 0.5$, then we have that

$$h^*(X) = sign\left(w^T \left(X - \frac{\mu_1 + \mu_2}{2}\right)\right)$$

where

$$w^T \left(\frac{\mu_1 + \mu_2}{2}\right) = -b$$

3 Fischer Discriminant

Let $X \in \mathbb{R}^d$ with p.d.f $f_X(x)$ such that $E(X) = \mu$ and $C_{ij} = E\left[(X_i - \mu_i)(X_j - \mu_j)\right]$. Then for any vector $u \in \mathbb{R}^d$, if $z = u^T X$ we have that

$$\begin{aligned} E(z) &= E(u^T X) \\ &= u^T E(X) \\ &= u^T \mu \\ \text{Var}(z) &= E\left(u^T X - E(z)\right)^2 \\ &= E\left(u^T X - u^T \mu\right)^2 \\ &= E[u^T (X - \mu)(X - \mu)^T u] \\ &= u^T C u \end{aligned}$$

We define for an arbitrary vector $w \in \mathbb{R}^d$,

$$\|w\| = \sqrt{w^T w} = \sqrt{\sum_i w_i^2}$$

Consider $C = \{\alpha w \mid \alpha \in \mathbb{R}\}$. Projection of any $v \in \mathbb{R}^d$ on C is the point in C which is closest to v. Thus, we have to find

$$\min_{\alpha \in \mathbb{R}} \|v - \alpha w\|^2 \geq 0 \quad v, w \in \mathbb{R}^d$$

For ease of notation, we define

$$\alpha^* = \frac{v^T w}{\|w\|^2}$$

Using this, we have

$$\begin{aligned} \|v - \alpha w\|^2 &= \|v\|^2 - 2\alpha v^T w + \alpha^2 \|w\|^2 \\ &= \|v\|^2 - 2\alpha\alpha^* \|w\|^2 + \alpha^2 \|w\|^2 \\ &= \|v\|^2 c + \|w\|^2 (\alpha - \alpha^*)^2 \\ &\geq \|v\|^2 - \alpha^2 \|w\|^2 \end{aligned}$$

Therefore, α^* minimizes $\|v - \alpha w\|^2$ and since $\|v\|^2 - \alpha^2 \|w\|^2 \geq 0$, we obtain

$$\|w\|^2 \|v\|^2 \geq (v^T w)^2$$

Thus, for any $v, w \in \mathbb{R}^d$, we have that $\|w\|^2 \|v\|^2 \geq (v^T w)^2$ with equality iff $w = tv$. Hence, for any $x \in \mathbb{R}^d$, $w^T x$ is the projection of x on w. We can reduce the dimensionality of the feature space from d dimensions to one dimension in the hope of a more manageable problem if we merely project the d-dimensional data onto a line. We now turn to the matter of finding the best

such direction w of the line, one we hope will enable accurate classification. A measure of the separation between the projected points is the difference of the sample means. Of course, to obtain good separation of the projected data we really want the difference between the means to be large relative to some measure of the standard deviations for each class.

Thus, after projection, we need $|w^T\mu_1 - w^T\mu_2|$ to be large as well as $w^T C_1 w$ and $w^T C_2 w$ to be small. Hence, we need to find

$$\max_w \frac{|w^T(\mu_1 - \mu_2)|^2}{w^T C_1 w + w^T C_2 w}$$

Note that if $B = C_1 + C_2$, then the denominator can be rewritten as $w^T Bw$. Similarly, we rewrite the numerator as $w^T Aw$ where $A = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$. It can be seen that both A and B are symmetric and positive definite matrices i.e., $A, B \in S_d^+$. Hence, we have

$$\max_{w\in\mathbb{R}^d} \frac{w^T Aw}{w^T Bw}$$

Using the property, that the square root of a positive-definite matrix is also positive definite, we express B as L^2 where $L \in S_d^+$ i.e., $L = B^{1/2}$. Let $u = B^{1/2}w$, then $w = B^{-1/2}u$. Thus, we get

$$\max_{u\in\mathbb{R}^d} \frac{u^T B^{-1/2} A B^{-1/2} u}{u^T u}$$

which takes the following form for some $E \in S_d$,

$$\max_{u\in\mathbb{R}^d} \frac{u^T Eu}{u^T u} = \lambda_{\max}(E) = \frac{u_o^T E u_o}{u_o^T u_o}$$

Here u_o is the eigenvector of $\lambda_{\max}(E)$ where $\lambda_{\max}$ is the largest eigenvalue. Thus,

$$\max_{u\in\mathbb{R}^d} \frac{u^T B^{-1/2} A B^{-1/2u}}{u^T u} = \lambda_{\max}(B^{-1/2} A B^{-1/2}) = \bar{\lambda}$$

is attained at

$$B^{-1/2} A B^{-1/2} u_o = \bar{\lambda} u_o$$

Using $w = B^{-1/2}u$,

$$w_o = \arg\max_{w\in\mathbb{R}^d} \frac{w^T Aw}{w^T Bw}$$

Since, $w_o = B^{-1/2}u_o$, we have

$$B^{-1}Aw_o = \bar{\lambda} w_o$$
$$\implies Aw_o = \bar{\lambda} B w_o$$

The last equation is well known as the Generalised Eigenvalue Problem. For Fisher discriminant,

$$A = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T, B = C_1 + C_2$$
$$\implies \bar{\lambda} w_o = (c_1 + C_2)^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w_o$$

Noting that $\bar{\lambda}(\mu_1 - \mu_2)^T w_o = (\mu_1 - \mu_2)^T (C_1 + C_2)^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w_o$ clearly yields that $\bar{\lambda} = (\mu_1 - \mu_2)^T (C_1 + C_2)^{-1}(\mu_1 - \mu_2)$ and $w_o = (C_1 + C_2)^{-1}(\mu_1 - \mu_2)$ since $(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ always lies in the direction of $(\mu_1 - \mu_2)$ and the scale factor for w_o is immaterial.

4 Perceptron Algorithm

McCulloch and Walter Pitts developed a neuron, which takes as input a d-dimensional vector $x_0, x_1, \ldots, x_d$ and returns a prediction y given by ($\mathbf{w}$ is also know as the weight vector)

$$y = sign\left(\sum_{j=0}^{d} w_j x_j\right)$$

For a perceptron in general, the input are $\mathcal{D} = \{(x^{(i)}, y^{(i)}) \mid x^{(i)} \in X,\ y^{(i)} \in \{-1, 1\}\}$ where $i \in [N]$. We shall now assume that there exists $\mathbf{w}^*$ such that

$$sign\left((\mathbf{w}^*)^T x^{(i)}\right) = y^{(i)} \quad , i \in [N]$$

$$\implies y^{(i)}\left((\mathbf{w}^*)^T x^{(i)}\right) > 0$$

Our task is to find the weight vector $\mathbf{w}$. We shall consider $\mathbf{w}^{(1)} = 0$ and obtain a formula for $\mathbf{w}^{(n+1)}$ in terms of $\mathbf{w}^{(n)}$. Suppose that $\mathbf{w}^{(n)}$ is the current estimate and consider an $(x^{(l)}, y^{(l)}) \in \mathcal{D}$ such that

$$sign\left((\mathbf{w}^{(n)})^T x^{(l)}\right) \neq y^{(l)}$$

$$\implies y^{(l)}\left((\mathbf{w}^{(n)})^T x^{(l)}\right) < 0$$

Hence, we obtain the following

$$\mathbf{w}^{(n+1)} = \begin{cases} \mathbf{w}^{(n)} + y^{(l)} x^{(l)} & \text{update} \\ \mathbf{w}^{(n)} & \text{otherwise} \end{cases}$$

Now, lets compute

$$\begin{aligned} \|\mathbf{w}^{(n+1)}\|^2 - \|\mathbf{w}^{(n)}\|^2 &= (\mathbf{w}^{(n+1)} - \mathbf{w}^{(n)})^T (\mathbf{w}^{(n+1)} - \mathbf{w}^{(n)}) \\ &= (y^{(l)} x^{(l)})^T (2\mathbf{w}^{(n)} + y^{(l)} x^{(l)}) \\ &= 2y^{(l)} (\mathbf{w}^{(n)})^T x^{(l)} + \|x^{(l)}\|^2 \\ &\leq \|x^{(l)}\|^2 \qquad\qquad \because y^{(l)} (\mathbf{w}^{(n)})^T x^{(l)} < 0 \end{aligned}$$

If we assume $\|x^{(i)}\| \leq R$ for $i \in [N]$, then $\|\mathbf{w}^{(n+1)}\|^2 - \|\mathbf{w}^{(n)}\|^2 \leq R^2$. Let us define a new quantity ν.

$$\nu = \min_i \frac{y^{(i)} (\mathbf{w}^*)^T x^{(i)}}{\|\mathbf{w}^*\|}$$

Then,

$$\therefore (\mathbf{w}^*)^T (\mathbf{w}^{(n+1)} - \mathbf{w}^{(n)}) = (\mathbf{w}^*)^T (y^{(l)} x^{(l)}) \geq \nu \|\mathbf{w}^*\|$$

Let M be the number of updates till there are no mistakes. Then after M updates,

$$\sum_{k=1}^{M} (\mathbf{w}^*)^T (\mathbf{w}^{(n+1)} - \mathbf{w}^{(n)}) \geq M\nu\|\mathbf{w}^*\|$$
$$\implies (\mathbf{w}^*)^T (\mathbf{w}^{(M+1)} - \mathbf{w}^{(1)}) \geq M\nu\|\mathbf{w}^*\|$$

If we use $\mathbf{w}^{(1)} = 0$, then we can rewrite the above equation as

$$(\mathbf{w}^*)^T \mathbf{w}^{(M+1)} \geq M\nu\|\mathbf{w}^*\|$$

Thereby, using Cauchy-Schwartz inequality, we get

$$M\nu \leq \|\mathbf{w}^{(M+1)}\|$$

Using $\|\mathbf{w}^{(n+1)}\|^2 - \|\mathbf{w}^{(n)}\|^2 \leq R^2$, we get

$$\sum_{k=1}^{M} \|\mathbf{w}^{(n+1)}\|^2 - \|\mathbf{w}^{(n)}\|^2 \leq MR^2$$
$$\implies \|\mathbf{w}^{(M+1)}\|^2 \leq MR^2$$
$$\therefore M^2\nu^2 \leq MR^2$$
$$\implies M \leq \frac{R^2}{\nu^2}$$

We thus obtain that, on a linearly separable dataset (i.e., a set $\mathcal{X}$ of vectors in $\mathbb{R}^d$ which can be classified into two sets $\mathcal{X}^+$ and $\mathcal{X}^-$ with $\mathcal{X}^+ = \{\mathbf{x} \in \mathcal{X} : \mathbf{w} \cdot \mathbf{x} > 0\}$ and $\mathcal{X}^- = \{\mathbf{x} \in \mathcal{X} : \mathbf{w} \cdot \mathbf{x} < 0\}$ for some $\mathbf{w} \in \mathbb{R}^d$), the perceptron algorithm terminates after making at most M updates, where M is given as

$$M \leq \frac{R^2\|\mathbf{w}^*\|^2}{\min_i (y^{(i)}(\mathbf{w}^*)^T x^{(i)})}$$

Lets consider $\mathcal{U} = \{i_1, \ldots, i_M\}$ i.e., the indices which are updated. Then if $\overline{\mathbf{w}}$ is defined as below

$$\overline{\mathbf{w}} = \sum_{k=1}^{M} y^{(i_k)} x^{(i_k)}$$

we have that

$$sign(\overline{\mathbf{w}}^T x^{(i)}) = y^{(i)} \quad , i \in \{1, \ldots, N\}$$

Let us denote the classifier returned by the perceptron algorithm by $h_{\mathcal{D}}^{(P)}$ where

$$h_{\mathcal{D}}^{(P)} = sign(\overline{\mathbf{w}}_D^T X) \quad , \quad \overline{\mathbf{w}}_D = \sum_{k=1}^{M} y^{(i_k)} x^{(i_k)}$$

$$\therefore 1_{\{h_{\mathcal{D}}^{(P)}(X^{(i)}) \neq Y^{(i)}\}} = 0 \because \forall\, i \in [N]\ sign(\overline{\mathbf{w}}^T x^{(i)}) = y^{(i)}$$

Then its risk (or generalization error) is given by (here $\mathcal{D}$ is a sample of size N)

$$P(h_{\mathcal{D}}^{(P)} \neq Y) = R(h_{\mathcal{D}}^{(P)})$$

We now try to find the expected generalization error by a classifier returned by perceptron algorithm acting on a linearly separable sample of N i.i.d draws from P, denoted henceforth as

$$E_{\mathcal{D}\sim P^{(N)}} R(h_{\mathcal{D}}^{(P)})$$

The hack we shall use is

- Consider $\overline{\mathcal{D}} \sim P^{(N+1)}$ which is a sample of size $N+1$
- Let $\overline{\mathcal{D}}$ be linearly separable
- Create $N+1$ datasets from $\overline{\mathcal{D}}$ by removing the i-th data point i.e.,

  $$\mathcal{D}^{(i)} = \overline{\mathcal{D}} \setminus \{x^{(i)}, y^{(i)}\} \quad , i = 1, \ldots, N+1$$

  Then, we have that
  - size of $\mathcal{D}^{(i)} = N$
  - $\mathcal{D}^{(i)}$ is linearly separable since $\overline{\mathcal{D}}$ is linearly separable

Now, consider an algorithm A acting on a sample $\mathcal{D}$ of size m that returns a classifier $h_{\mathcal{D}}^{(A)}$. Then the leave one out error (LOO) of A on $\mathcal{D}$ is defined as

$$\overline{R}_{\mathcal{D}}^{\text{LOO}}(A) = \frac{1}{m} \sum_{i=1}^{m} 1_{\{h_{\mathcal{D}^{(i)}}^{(A)}(X^{(i)}) \neq Y^{(i)}\}}$$

Lets now find the expected LOO error on a random sample of size m.

$$E_{\mathcal{D}\sim P^{(m)}} \overline{R}_{\mathcal{D}}^{\text{LOO}}(A) = \frac{1}{m} E_{\mathcal{D}\sim P^{(m)}} \sum_{i=1}^{m} 1_{\{h_{\mathcal{D}^{(i)}}^{(A)}(X^{(i)}) \neq Y^{(i)}\}}$$

Since, the points of $\mathcal{D}^{(m+1)}$ are drawn in an i.i.d fashion, the following expectation

$$E_{\mathcal{D}\sim P^{(m)}} 1_{\{h_{\mathcal{D}^{(i)}}^{(A)}(X^{(i)}) \neq Y^{(i)}\}}$$

does not depend on the choice of $i \in [m]$. In other words, we have that

$$E_{\mathcal{D}\sim P^{(m)}} = E_{\mathcal{D}^{(i)}\sim P^{(m-1)}} E_{X^{(i)},Y^{(i)}\sim P}$$

Now,

$$E_{X^{(i)},Y^{(i)}\sim P} 1_{\{h_{\mathcal{D}^{(i)}}^{(A)}(X^{(i)}) \neq Y^{(i)}\}} = P(h_{\mathcal{D}^{(i)}}^{(A)}(X^{(i)}) \neq Y^{(i)}) = R(h_{\mathcal{D}^{(i)}}^{(A)})$$

Thus,

$$\begin{aligned} E_{\mathcal{D}\sim P^{(m)}} \overline{R}_{\mathcal{D}}^{\text{LOO}}(A) &= \frac{1}{m} \sum_{i=1}^{m} E_{\mathcal{D}^{(i)}\sim P^{(m-1)}} R(h_{\mathcal{D}^{(i)}}^{(A)}) \\ &= E_{\mathcal{D}\sim P^{(m-1)}} R(h_{\mathcal{D}}^{(A)}) \end{aligned}$$

We have thus shown that average generalization error of a classifier derived by an algorithm A acting on a sample of size $m-1$ is equal to the expected LOO error of A on sample of size m i.e.,

$$E_{\mathcal{D}\sim P^{(m)}}\overline{R}_{\mathcal{D}}^{\text{LOO}}(A) = E_{\mathcal{D}\sim P^{(m-1)}}R(h_{\mathcal{D}}^{(A)})$$

Now, returning to the perceptron, we see that for $j \in [N]$

- if $j \notin \mathcal{U}$ then

$$\begin{aligned} h_{\mathcal{D}^{(j)}}^{(P)}(X) &= h_{\mathcal{D}}^{(P)}(X) \\ \implies 1_{\{h_{\mathcal{D}^{(j)}}^{(P)}(X^{(i)})\neq Y^{(i)}\}} &= 0 \end{aligned}$$

- if $j \in \mathcal{U}$ then

$$1_{\{h_{\mathcal{D}^{(j)}}^{(P)}(X^{(i)})\neq Y^{(i)}\}} \leq 1$$

Thus,

$$\begin{aligned} \overline{R}_{\mathcal{D}}^{\text{LOO}}(P) &= \frac{1}{N+1}\sum_{j=1}^{N+1} 1_{\{h_{\mathcal{D}^{(j)}}^{(P)}(X^{(i)})\neq Y^{(i)}\}} \\ &\leq \frac{1}{N+1}\left(\sum_{j\notin\mathcal{U}} 0 + \sum_{j\in\mathcal{U}} 1\right) \\ &= \frac{M}{N+1} \end{aligned}$$

It now follows that,

$$\begin{aligned} E_{\mathcal{D}\sim P^{(N)}}R(h_{\mathcal{D}}^{(P)}) &= E_{\mathcal{D}\sim P^{(N+1)}}\overline{R}_{\mathcal{D}}^{\text{LOO}}(P) \\ &\leq E_{\mathcal{D}\sim P^{(N+1)}}\frac{\min(M(\mathcal{D}), R^2(\mathcal{D})/\nu^2(\mathcal{D}))}{N+1} \end{aligned}$$

5 Primer on Convex Optimization and KKT conditions

§5.1 Convex Optimization Problem

A set $C \subseteq \mathbb{R}^d$ is said to be convex if for all $\mathbf{x}^{(1)} \in C$ and $x^{(2)} \in C$ then $\alpha x^{(1)} + (1-\alpha)x^{(2)} \in C$ for all $0 < \alpha < 1$. Thus, $[a, b] = \{x \mid a \leq x \leq b\}$ is a convex set. A function $f : [a, b] \mapsto \mathbb{R}$ is said to be convex if for all $x^{(1)} \neq x^{(2)} \in [a, b]$, we have that

$$f((1-\alpha)x^{(1)} + \alpha x^{(2)}) \leq (1-\alpha)f(x^{(1)}) + \alpha f(x^{(2)})$$

Minimizing a convex function over a closed convex set is called a Convex Optimization Problem (CVO). Now, consider the constraint set given as

$$C = \{x \in \mathbb{R}^d \mid a_i^T x \geq b_i \quad i = 1, \ldots, m\}$$

where $a \in \mathbb{R}^d$, $b \in \mathbb{R}$ and x is the decision variable. Given an objective function $f(x)$ such that $f : C \mapsto \mathbb{R}$, we are required to find

$$\begin{aligned} f^*(x) &= \min_{x \in C \subseteq \mathbb{R}^d} f(x) \quad \implies f(x) \geq f^*(x) \, \forall \, x \in C \\ x^* &= \operatorname*{argmin}_{x \in C \subseteq \mathbb{R}^d} f(x) \quad \implies f(x^*) \leq f(x) \, \forall \, x \in C \end{aligned}$$

If

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

then the gradient of f at $x \in C$ is the vector in $\mathbb{R}^d$ denoted by $\nabla f(x)$ and defined as

$$\nabla f(x) = \begin{bmatrix} \dfrac{\partial f}{\partial x_1}(x) \\ \vdots \\ \dfrac{\partial f}{\partial x_d}(x) \end{bmatrix}$$

Note that here the $(\partial f / \partial x_i)$'s are continuous. Suppose that $f \in C^1$ is defined over a convex set $C \subseteq \mathbb{R}^d$, then f is convex over C if and only if, for all $x, y \in C$

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

Similarly, if $f \in C^2$ is defined over a convex set $C \subseteq \mathbb{R}^d$, then f is convex over C if and only if, for all $x \in C$

$$H(x) \geq 0 \quad \text{where } (H(x))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x)$$

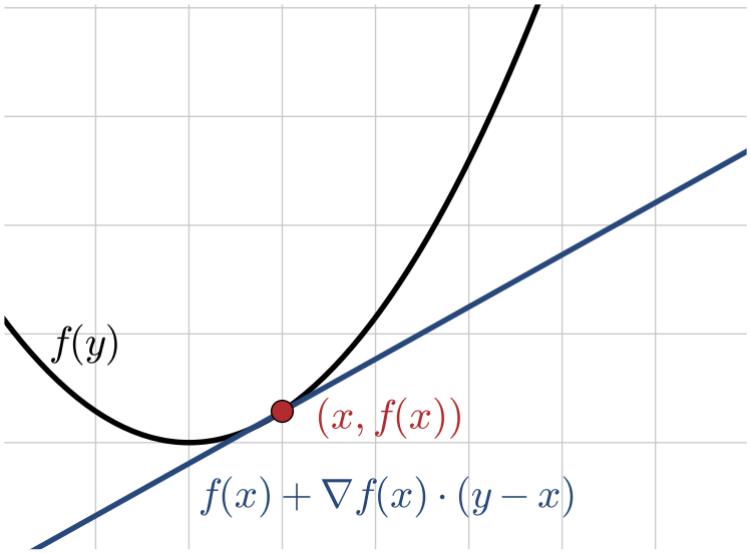


Illustration of the first-order property satisfied by all convex functions

$H(x)$ is also called Hessian of f at $x \in C$ which is a vector in $\mathbb{R}^{d\times d}$. Observe that $H(x)$ is symmetric and positive semi-definite.

Moving forward given these preliminaries, now consider a convex function $g : \mathbb{R}^d \mapsto \mathbb{R}$. Then $\{x \mid g(x) \leq t\}$ is a convex set for all $t \in \mathbb{R}$ i.e.,

$$g((1-\alpha)x^{(1)} + \alpha x^{(2)}) \leq (1-\alpha)g(x^{(1)}) + \alpha g(x^{(2)}) \leq t$$

§5.2 K.K.T Conditions

Given convex function f, f_i, consider the following problem P where we have to find

$$\min_{x\in\mathbb{R}^d} f(x)$$

subject to

$$f_i(x) \leq 0 \quad i = 1, \ldots, m$$
$$a_j^T x = b_j \quad j = 1, \ldots, n$$

We introduce Lagrange variables $\lambda_i \geq 0$, $i \in [m]$ and $\mu_j \geq 0$, $j \in [n]$ associated to the constraints and denote by λ the vector $(\lambda_1, \ldots, \lambda_m)^T$ as well as by μ the vector $(\mu_1, \ldots, \mu_n)^T$. The Lagrangian can then be defined for all $x \in \mathbb{R}^d$ by

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{n} \mu_j (a_j^T x - b_j)$$

If for any x^*, there exists λ^*, μ^* such that (x^*, λ^*, μ^*) satisfy the following three conditions, then x^* is known to be a KKT (Karush-Kuhn-Tucker) point of P.

- $$\nabla_x \mathcal{L} = \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{j=1}^{n} \mu_j a_j = 0$$
- $$\lambda_i f_i(x) = 0 \quad \lambda_i \geq 0$$
- $$f_i(x) \leq 0 \quad i = 1, \ldots, m$$
$$a_j^T x = b_j \quad j = 1, \ldots, n$$

***If x^* is a K.K.T point, then it is a global minimum of* P**

Lets consider anther example. P is now to find,

$$\min_{x \in \mathbb{R}^d} \frac{1}{2}||x - z||^2$$

subject to

$$w^T x + b = 0$$

Then, the Lagrangian is given as

$$\mathcal{L}(x, \mu) = \frac{1}{2}||x - z||^2 + \mu(w^T x + b)$$

Then, since $\nabla_x \mathcal{L} = x - z + \mu w = 0$, we get that

$$x = z - \mu w$$

Substituting this in $w^T x + b = 0$, we get $w^T z - \mu||w||^2 + b = 0$, which implies that

$$\mu^* = \frac{w^T z + b}{||w||^2} \quad , \quad x^* = z - \mu^* w$$

With these values, we get that

$$\begin{aligned} ||x^* - z|| &= ||\mu^* w|| \\ &= |\mu^*| \cdot ||w|| \\ &= \frac{|w^T z + b|}{||w||} \end{aligned}$$

This shows that x^* is a K.K.T point of P. Also, it can be seen that x^* is the global minimum of P since the shortest distance of z from $w^T x + b = 0$ is indeed $|w^T z + b|/||w||$.

6 Large Margin Classification

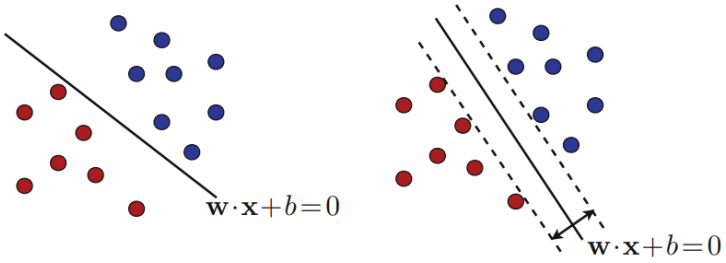
Lets recollect the setting of the Perceptron Algorithm. We are given an input space $X \subseteq \mathbb{R}^d$ and the label space $Y = \{-1, 1\}$. The binary classification task is formulated as follows – given input $\mathcal{D}$ of size N i.e., $\mathcal{D} = \{(x^{(i)}, y^{(i)}) \mid x^{(i)} \in X, y^{(i)} \in \{-1, 1\}\}$ where $i \in [N]$ we have to find a binary classifier $h : X \mapsto Y$ with minimal risk (or generalisation error). A natural classifier with relatively small complexity is that of linear classifiers, or hyperplanes, which we have seen previously i.e., $\{x \mapsto sign(w^T x + b) \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$.

§6.1 Separable Case

In this section, we assume that $\mathcal{D}$ can be linearly separated that is, we assume the existence of a hyperplane that perfectly separates the training sample into two populations of positively and negatively labeled points, as illustrated below. This is equivalent to the existence of $(w, b) \in (\mathbb{R}^N - \{0\}) \times \mathbb{R}$ such that

$$\forall\, i \in [N], \quad y^{(i)}(w^T x^{(i)} + b) > 0$$

However, as can be seen below, there are then infinitely many such separating hyperplanes. Which hyperplane should we then select ?



Let us define a new quantity $\gamma(w)$ as follows,

$$\gamma(w) = \min_i \frac{y^{(i)}(w^T x^{(i)} + b)}{||w||}$$

We would like to find $\gamma = \max_{w,b} \gamma(w)$. We observe that if (w^*, b^*) solves the problem, then so does (sw^*, sb^*) where $s > 0$. Thus, we choose s such that $\gamma||w|| = 1$ i.e., $\min_i y^{(i)}(w^T x^{(i)} + b) = 1$. Hence, the problem now reduces to finding

$$\max_{\substack{w,b \\ \min_{i\in[m]} y^{(i)}(w^T x^{(i)}+b)=1}} \frac{1}{||w||} = \max_{\substack{w,b \\ \forall i\in[m]\,,\, y^{(i)}(w^T x^{(i)}+b)\geq 1}} \frac{1}{||w||}$$

Equivalent Problem : Since maximizing $1/\|w\|$ is equivalent to minimizing $(1/2)\|w\|^2$, it is equivalent to the following convex optimization problem

$$\min_{w,b} \ \frac{1}{2}\|w\|^2$$

subject to

$$y^{(i)}(w^T \cdot x^{(i)} + b) \geq 1 \, , \, \forall \, i = 1, \ldots, N$$

We introduce Lagrange variables $\lambda_i \geq 0, i \in [N]$ associated to the N constraints and denote by λ the vector $(\lambda_1, \ldots, \lambda_m)^\top$. The Lagrangian can then be defined for all $w \in \mathbb{R}^N, b \in \mathbb{R}$ and $\lambda \in \mathbb{R}^N_+$ by

$$\mathcal{L}(w, b, \lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \lambda_i [y^{(i)}(w^T x^{(i)} + b) - 1]$$

Then, we have from K.K.T conditions that

$$\begin{aligned} \nabla_w \mathcal{L} = w - \sum_{i=1}^{N} \lambda_i y^{(i)} x^{(i)} \quad &= 0 \\ \nabla_b \mathcal{L} = - \sum_{i=1}^{N} \lambda_i y^{(i)} \quad &= 0 \end{aligned}$$

From K.K.T conditions, we get that $\forall \, i = 1, \ldots, N$

$$\begin{aligned} \sum_{i=1}^{N} \lambda_i y^{(i)} x^{(i)} &= w \\ \lambda_i [y^{(i)}(w^T x^{(i)} + b) - 1] &= 0 \\ y^{(i)}(w^T \cdot x^{(i)} + b) &\geq 1 \end{aligned}$$

Since $\lambda_i \geq 0$, we obtain two cases

- If $\lambda_i > 0$, then $y^{(i)}(w^T x^{(i)} + b) = 1$
- If $y^{(i)}(w^T x^{(i)} + b) > 1$, then $\lambda_i = 0$

Also since $y^{(i)} \in \{-1, 1\}$ and $y^{(i)}(w^T x^{(i)} + b) \geq 1$, we again obtain two cases

- If $(w^T x^{(i)} + b) \geq 1$, then $y^{(i)} = 1$
- If $(w^T x^{(i)} + b) \leq -1$, then $y^{(i)} = -1$

A support vector is a vector $x^{(i)}$ which occurs in the expansion if and only if $\lambda_i > 0$ in which case $y^{(i)}(w^T x^{(i)} + b) = 1$, thereby lying on the hyperplanes $w^T x^{(i)} + b = \pm 1$.

§6.2 Non-Separable Case

Consider the primal problem P given in the K.K.T section. Then, the corresponding dual can be given as,

$$\max_{x,\lambda,\mu} \mathcal{L}(x,\lambda,\mu)$$

subject to

$$\nabla_x \mathcal{L}(x,\lambda,\mu) = 0$$
$$\lambda \geq 0$$

It is evident that the K.K.T point $(x^*,\ \lambda^*,\ \mu^*)$ solves the above dual, also known as Wolfe Dual i.e., $\nabla_x \mathcal{L}(x^*,\ \lambda^*,\ \mu^*) = 0$ and $\lambda^* \geq 0$. Hence, by strong duality we conclude that

$$\mathcal{L}(x^*,\ \lambda^*,\ \mu^*) = f(x^*)$$

Since we know that f, f_i are convex functions, hence for any $x \in \mathbb{R}^d$, we have

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x)$$
$$f_i(x^*) \geq f_i(x) + \nabla f_i(x)^T (x^* - x)$$
$$a_j^T x^* - b = a_j^T x - b + a_j^T (x - x^*)$$

Using these, we have for any $\lambda_i \geq 0$

$$\begin{aligned}
f(x^*) &= \mathcal{L}(x^*,\ \lambda^*,\ \mu^*) \\
&\geq f(x^*) + \sum_{i=1}^{m} \lambda_i f_i(x^*) + \sum_{j=1}^{n} \mu_j (a_j^T x^* - b_j) \\
&\geq f(x) + \nabla f(x)^T (x^* - x) + \sum_{i=1}^{m} \lambda_i \left(f_i(x) + \nabla f_i(x)^T (x^* - x) \right) + \sum_{j=1}^{n} \mu_j \left(a_j^T x - b \right) + \sum_{j=1}^{n} \mu_j a_j^T (x - x^*) \\
&\geq f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{n} \mu_j \left(a_j^T x - b \right) + (x^* - x)^T \left[\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{j=1}^{n} \mu_j a_j \right] \\
&= \mathcal{L}(x,\lambda,\mu) + (x^* - x)^T \nabla_x \mathcal{L}(x,\lambda,\mu) \\
&\geq \mathcal{L}(x,\lambda,\mu)
\end{aligned}$$

Lets see the Wolfe Dual for the SVM problem which is to find

$$\max_{w,b,\lambda} \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \lambda_i [y^{(i)}(w^T x^{(i)} + b) - 1]$$

subject to

$$\sum_{i=1}^{N} \lambda_i y^{(i)} x^{(i)} = w$$
$$\sum_{i=1}^{N} \lambda_i y^{(i)} = 0$$

Now eliminating w_i, b, we get

$$\max_{\lambda \geq 0} \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \Big\| \sum_{i=1}^{N} \lambda_i y^{(i)} x^{(i)} \Big\|^2 \quad \text{subject to } \sum_{}^{N} \lambda_i y^{(i)} = 0$$

which after expanding the norm becomes,

$$\max_{\lambda \geq 0} \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y^{(i)} y^{(j)} (x^{(i)})^T x^j \quad \text{subject to } \sum_{}^{N} \lambda_i y^{(i)} = 0$$

7 Kernel Methods

Now, lets solve SVM in the feasible space where we shall consider an (embedding) map $\phi(x)$ instead of x which aims to convert the nonlinear relations into linear ones. Now, the dual becomes

$$\max_{\lambda \geq 0} \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y^{(i)} y^{(j)} \phi(x^{(i)})^T \phi(x^j) \quad \text{subject to } \sum_{i=1}^{N} \lambda_i y^{(i)} = 0$$

and our classifier is given as

$$h(x) = sign\left(\sum_{i=1}^{N} \lambda_i y^{(i)} \phi(x^{(i)})^T \phi(x) + b \right)$$

For example, if we consider

$$\phi(x) = \begin{bmatrix} 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ \sqrt{2}x_1 x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix} \quad \phi(\tilde{x}) = \begin{bmatrix} 1 \\ \sqrt{2}\tilde{x}_1 \\ \sqrt{2}\tilde{x}_2 \\ \sqrt{2}\tilde{x}_1 \tilde{x}_2 \\ \tilde{x}_1^2 \\ \tilde{x}_2^2 \end{bmatrix}$$

then

$$\begin{aligned} \phi(x)^T \phi(\tilde{x}) &= 1 + 2x_1\tilde{x}_1 + 2x_2\tilde{x}_2 + 2x_1x_2\tilde{x}_1\tilde{x}_2 + x_1^2\tilde{x}_1^2 + x_2^2\tilde{x}_2^2 \\ &= (1 + x_1\tilde{x}_1 + x_2\tilde{x}_2)^2 \\ &= (1 + x^T\tilde{x})^2 \\ &= K(x, \tilde{x}) \end{aligned}$$

where $x = [x_1\ x_2]^T$, $\tilde{x} = [\tilde{x}_1\ \tilde{x}_2]^T$ and $K : X \times X \mapsto \mathbb{R}$ is called a kernel function if

- K is symmetric
- K is positive semi-definite

From now, we shall use the following notation, where for every $n \in \mathbb{N}$ and for every $D_n = \{x^{(i)} \in X \mid i \in [n]\}$, we have that $K_{ij} = K(x^{(i)}, x^{(j)})$ and $K \geq 0$. For example, consider $K(x, z) = x^T z = K(z, x)$, hence it is symmetric and $K_{ij} = K(x^{(i)}, x^{(j)}) = (x^{(i)})^T x^{(j)}$. It is positive semi-definite as well, since for all u, we observe that

$$\begin{aligned} u^T K u &= \sum_i \sum_j u_i u_j (x^{(i)})^T x^{(j)} \\ &= \left(\sum_i u_i x^{(i)} \right)^T \left(\sum_j u_j x^{(j)} \right) \\ &= \| \sum_i u_i x^{(i)} \|^2 \geq 0 \end{aligned}$$

§7.1 Mercer's Theorem

According to Mercer's Theorem, let $X \subseteq \mathbb{R}^d$ be a compact set and let $K : X \times X \mapsto \mathbb{R}$ be a continuous symmetric function. Then for every $f \in L_2(X)$ (sqaure-integrable functions), we have that

$$\int_{X\times X} K(x,z)f(x)f(z)dxdz \geq 0$$

is true if and only if there exists $\phi_i \in L_2(X)$ satisfying $\int \phi_i(x)\phi_j(x)dx = 0$ such that

$$K(x,z) = \sum_{i=1}^{\infty} \phi_i(x)\phi_i(z)$$

§7.2 Normalized Kernel

Let $K(x,z)$ be a valid kernel, then we shall that $\tilde{K}(x,z)$ defined as follows, is also a valid kernel

$$\tilde{K}(x,z) = \frac{K(x,z)}{\sqrt{K(x,x)}\sqrt{K(z,z)}}$$

We can see $\tilde{K}(x,z)$ is symmetric since,

$$\tilde{K}(x,z) = \frac{\phi(x)^T\phi(z)}{||\phi(x)|| \cdot ||\phi(z)||} = \tilde{K}(z,x)$$

It is also positive semi-definite as well since, for all $u \in \mathbb{R}^N$ if $\tilde{K}_{ij} = \tilde{K}(x^{(i)}, x^{(j)})$

$$\begin{aligned} u^T\tilde{K}u &= \sum_i \sum_j u_i u_j \frac{\phi(x^{(i)})^T\phi(x^{(j)})}{||\phi(x^{(i)})|| \cdot ||\phi(x^j)||} \\ &= \Big|\Big| \sum_i u_i \frac{\phi(x^{(i)})}{||\phi(x^{(i)})||} \Big|\Big|^2 \\ &\geq 0 \end{aligned}$$

$\tilde{K}$ is called normalized kernel since $\tilde{K}(x,x) = 1$.

§7.3 Gaussian Kernel

We see that if K_1 and K_2 are two kernel functions, then for any $\alpha, \beta \geq 0$, $K = \alpha K_1 + \beta K_2$ is also a valid kernel function and so is $K(x,z) = K_1(x,z)K_2(x,z)$. Also $K(x,z) = \exp(x^T, z)$ is a kernel function since the exponential function can be arbitrarily closely approximated by polynomials with positive coefficients and hence is a limit of kernels and the finitely positive semi-definiteness property is closed under taking point wise limits i.e.,

$$K(x,z) = \lim_{m\to\infty} \sum_{i=0}^{m} \frac{1}{i!}(x^Tz)^i = e^{x^Tz}$$

We now normalize K to obtain the Gaussian Kernel $\tilde{K}$

$$\begin{aligned}\tilde{K}(x,z) &= \frac{e^{x^T z}}{e^{x^T x/2} \cdot e^{z^T z/z}} \\ &= e^{-\|x-z\|^2/2}\end{aligned}$$

§7.4 Covariance Matrix

Observe that since, $K_{ij} = K((x^{(i)}, x^{(j)}))$ is symmetric and positive semi-definite, it can also be regarded as a covariance matrix. Thus, its sensible to consider $A \sim N(0, K_1)$ and $B \sim N(0, K_2)$ where K_1, K_2 are two kernel functions and $A, B \in \mathbb{R}^n$.

II

Post Midsem

8 Soft Margin SVM

9 Regression

10 Bias-Variance Decomposition

11 Maximum Likelihood estimation

12 EM Algorithm