Q-learning + Epsilon Greedy:

Q-learning algorithm: Find for all $s \in S$ and $a \in A$ a function $Q(s,a)$, which gives a good approximation for $Q^*(s,a)$

1. Start with a random values for $Q(s,a)$ (eg all zero)
2. Choose a starting state $s_0 \in S$
3. Look up the current best action in that state. i.e ...

$$a_0 = \underset{a \in A}{\operatorname{argmax}}\ Q(s,a)$$

OR choose a Random action $a_0 \in A$ with probability $\varepsilon \in [0,1]$ (Epsilon greedy algo)

4. Apply this action and get a new state s_1 and reward

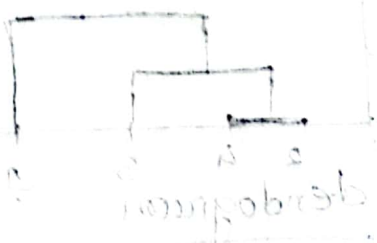
$$r_0 = R(s_0, a, s_1)$$

5. Update the value $Q(s_0, a_0)$ as follow (Bellman Equation)

$$Q(s_0,a_0) = (1-\alpha)\,Q(s_0,a_0) + \alpha\left(r_0 + \gamma \max_{a \in S} Q(s_1,a)\right)$$

Here, $\alpha \in [0,1]$ is the 'learning rate'

6. If s_1 is not the terminal state repeat with step 3.



AdaBoost Algo

1. For $i = 1$ to N Initialize that data weight $\omega_1^{(i)} = \frac{1}{N}$

2. For $t = 1$ to T

a) Find a classifier $h_t(x)$ by minimizing the weighted error function:

$$J_t = \sum_{i=1}^{N} \omega_t^{(i)} \times L\left(y^{(i)} \neq h_t\left(x^{(i)}\right)\right)$$

miss classify এর weight বাড়ছি।

b) Find the weighted error of $h_t(x)$:

$$\varepsilon_t = \frac{\sum_{i=1}^{N} \omega_t^{(i)} \times L\left(y^{(i)} \neq h_t(x^{(i)})\right)}{\sum_{i=1}^{N} \omega_t^{(i)}}$$

error

and a new component is assigned votes based on its

$$\alpha_t = \ln\left((1-\varepsilon_t)/\varepsilon_t\right)$$

we will asign some points /value to the Classifier. Between "a good classifier, that misclassify" and "bad classifier that misclassify", The first one is more significant or bad so we asign big α value to it to maximize its priority than others. And voting time, that classifier will be choosen whose α value is highest.

For weak classifier, $\varepsilon_t = 0.5$

$$\ln\left(\frac{1-0.5}{0.5}\right) = 0$$

$\therefore \alpha = 0$ [voting power is reduced]

Strong classifier, $\varepsilon_t = 0.4$

$$\ln\left(\frac{1-0.4}{0.4}\right) = 0.405$$

c) The normalized weights are updated:

$$\omega_{t+1}^{(i)} = \omega_t^{(t)} \underbrace{e^{\alpha_t}}_{\downarrow} I\left(y^{(i)} \neq h_t\left(x^{(i)}\right)\right)$$

e was taken to reduce error exponentially

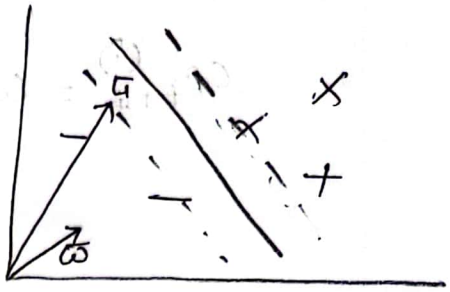
misclassify হলে $\alpha = 0$ $e^{\alpha} = 1$.

↑

strong

3. Combined classifier $\hat{y} = \text{sign}\left(H_T(x)\right)$ where $H_T(x) = \sum_{t=1}^{m} \alpha_t h_t(x)$

SVM



Proof: $\bar{u}$ → new value (unknown point)

$\vec{w} \cdot \bar{u} + b \geq 0$ then + [decision Rule]

→ Put additional constraints to calculate ω and b;

$\bar{\omega} \cdot \bar{x}_+ + b \geq 1$ ..... +Ve

$\vec{\omega} \cdot \bar{x}_- + b \leq -1$ ..... -Ve

→ Carrying two equation is pain.

→ Introduce new variable to make math convinient.

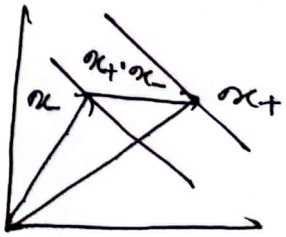
y_i such that $y_i = +$ for positive sample

$y_i = -$ for negative "

→ $y_i (\vec{\omega} \cdot \bar{x}_i + b) \geq 1$

→ $y_i (\bar{\omega} \cdot \bar{x}_i + b) - 1 \geq 0$

$\boxed{y_i (\bar{\omega} \cdot \bar{x}_i + b) - 1 = 0}$ [for x_i in the gutter/margi



Here, $Wx^+ + b = +1$

$\Rightarrow Wx^+ + b - 1 = 0$ ——(i)

$Wx^- + b + 1 = 0$ ——(ii)

$Wx^+ + b - 1 = Wx^- + b + 1$

$\Rightarrow W(x^+ - x^-) = 2$ —— (iii)

$\Rightarrow (x^+ - x^- = \frac{2}{\|w\|}$

$\Rightarrow m =$ ..

From figure $m = (x^+ - x^-)$

we need a direction so multiply by vector.

$$m = (x^+ - x^-)\frac{w}{\|w\|}$$

$$m = \frac{2}{\|w\|} \quad [\text{from equ (iii)}]$$

Maximize the margin:

$$\max \frac{2}{\|w\|} \sim \max \frac{1}{\|w\|} \sim \min \|w\| \sim \min \frac{\|w\|^2}{2}$$

minimize $\frac{\|w\|^2}{2}$ with respect to $y_i(\bar{w}\cdot\bar{x_i} + b) \geq 1)$

যেহেতু এখানে একটা শর্ত দেয়া আছে তাই G.D use না করে Lagrage Multiplier use হয়। Optimization problem-এ শর্ত দেয়া থাকলেই LM use হয়।

$$L = \frac{1}{2}\|w^2\| - \sum \alpha_i \left[y_i(\bar{w}\cdot\bar{x_i} + b) - 1\right] \quad \text{— (i)}$$

$$\frac{\partial L}{\partial w} = \bar{w} - \sum \alpha_i y_i x_i = 0$$

$$\Rightarrow \vec{w} = \sum \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = -\sum \alpha_i y_i = 0$$

$$\Rightarrow \sum \alpha_i y_i = 0.$$

Put the value of $\bar{w}$ in L eq(i)

$$L = \frac{1}{2}\left(\sum \alpha_i y_i x_i\right)\left(\sum \alpha_j y_j x_j\right) - \sum \alpha_i y_i x_i \left(\sum \alpha_j y_j x_j\right) - \sum \alpha_i y_i b + \sum \alpha_i$$

$$= \sum \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j$$

→ parameters are multiple to each other in L so its a quadratic optimization problem.

— Now use qp solver to find α

— all of the math depends on dot product. $(x_i \cdot x_j)$

Decision Rule: → dual optimization পদ্ধতি solve করেছি

$$y_i(\bar{\omega} x_i + b) \ldots$$

$$\sum \alpha_i y_i x_i \bar{u} + b \geq 0$$

Bayesian Decision theory

Conditional Probability, $P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)}$

$= \frac{\text{likehood} \times \text{prior}}{\text{evidence}}$.

Q. In Bayes parameter estimation for $\theta = \{\mu\}$, the estimate is calculated from the following equation:

$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} * \hat{x}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} * \mu_0$$

Show that the Bayes estimate of θ is same as that of MLE method when large number of samples is taken.

Ans: Given, μ_n is the MAP estimate of μ

$n \rightarrow$ number of sample

$\sigma^2 \rightarrow$ variance of the sample mean $\bar{x}$

$\sigma_0^2 \rightarrow$ Variance Associated with prior distribution of μ

$\bar{x} \rightarrow$ sample mean.

$\mu_0 \rightarrow$ mean associated with the prior distribution of μ

MLE for μ, assuming a Gaussian distribution, is simply the sample mean $\bar{x}$, as MLE doesnot account for prior distribution and maximizes the likelihood function given the observe data.

Now for MAPs equation,

(1) The term $\frac{n\sigma^2}{n\sigma_0^2 + \sigma^2}\bar{x}$ will dominate as n becomes very large, because $n\sigma_0^2$ will grow faster than σ^2, making the fraction approach 1.

(2) There term $\frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$ will diminish as n grows, because the denominator will grow much larger than the numerator, making the fraction approach to 0.

Therefore, for $n \to \infty$, the map estimate μ_n, converges to the sample mean $\bar{x}$, which is the MLE estimate for μ. So it concludes with large n, the prior becomes less influential and MAP estimate effectively becomes the MLE estimate.

#Define log-likelihood function and use Maximum Likelihood Estimation method for estimating the unknown parameter $\theta=(\mu,\sigma^2)$ of a univariate Gaussian distribution/ also $\theta=(\mu)$

⇒ For $\theta=(\mu)$ only mean is unknown.

$$P(x_k|\theta) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x_k-\theta)^t \Sigma^{-1}(x_k-\theta)\right]$$

We know, its easier to maximum if we add ln;

$$\ln P(x_k|\theta) = \ln \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x_k-\theta)^t \Sigma^{-1}(x_k-\theta)\right]$$

$$= -\frac{1}{2}(x_k-\theta)^t \Sigma^{-1}(x_k-\theta) - \frac{1}{2}\left[\ln(2\pi)^d|\Sigma|\right]$$

Now, $\nabla_\theta(\ln P(x_k|\theta)) = 0$

$$\Rightarrow \frac{\partial}{\partial\theta}\left\{-\frac{1}{2}\left[\ln(2\pi)^d|\Sigma|\right] - \frac{1}{2}(x_k-\theta)^t\Sigma^{-1}(x_k-\theta)\right\} = 0$$

$$\Rightarrow -\frac{1}{2}(x_k-\theta)\Sigma^{-1} = 0.$$

$$\Rightarrow \Sigma^{-1}(x_k-\theta) = 0$$

$$\Rightarrow \sum_{k=1}^{n} \Sigma^{-1}(x_k-\hat{\theta}) = 0$$

$$\Rightarrow \sum_{k=1}^{n} (x_k-\hat{\theta}) = 0$$

$$\Rightarrow \sum_{k=1}^{n} x_k - \sum_{k=1}^{n}\hat{\theta} = 0$$

$$\Rightarrow n\hat{\theta} = \sum_{k=1}^{n} x_k$$

$$\hat{\theta} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

For unknown mean and variance

Let $\theta = [\mu, \sigma^2]$ $\quad \theta_1 = \mu \quad \theta_2 = \sigma^2$

$$\ln(P(x_k|\theta)) = \frac{-1}{2} \ln[(2\pi)\theta_2] - \frac{1}{2}(x_k - \theta_1)^t \theta_2^{-1} (x_k - \theta_1)$$

$\nabla_\theta (\ln P(x_k|\theta))$:

with respect to $\theta_1 = \frac{1}{\theta_2}(x_k - \theta_1)$ ——— (i)

" " " $\theta_2 = \frac{-1}{2\theta_2} + \frac{(x-\theta_1)^2}{2\theta_2^2}$

From (i)

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0$$

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$

From (ii)

$$\sum_{k=1}^{n} -\frac{1}{2\hat{\theta}_2} + \frac{(x-\hat{\theta}_1)^2}{2\hat{\theta}_2^2} = 0$$

$$\Rightarrow \sum_{k=1}^{n} (x-\hat{\theta}_1)^2 = \sum_{k=1}^{n} \frac{2\hat{\theta}_2^2}{2\hat{\theta}_2}$$

$$\Rightarrow \sum_{k=1}^{n} (x-\hat{\theta}_1)^2 = \sum_{k=1}^{n} \hat{\theta}_2$$

$$\hat{\theta}_1 = \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k .$$

Derive a decision rule that minimizes the overall risks for a two-class problem.

Lets,

$\alpha_1 \rightarrow$ action $\rightarrow$ model predict feature to go to class ω_1

$\alpha_2 \rightarrow$ action $\rightarrow$ " " " " " " " ω_2

$\lambda_{ij} = \lambda(\alpha_i | \omega_j)$ predicted i class but actual j class.

↑ Loss function.

Suppose we observe x and take action α_i. The Conditional risk (or expected loss) with taking action α_i is,

$$R(\alpha_i | x) = \sum_{j=1}^{c} \lambda(\alpha_i | \omega_j) P(\omega_j | x)$$

Now For two class,

$$R(\alpha_1 | x) = \lambda_{11} P(\omega_1 | x) + \lambda_{12} P(\omega_2 | x)$$

$$R(\alpha_2 | x) = \lambda_{22} P(\omega_2 | x) + \lambda_{21} P(\omega_1 | x)$$

If $R(\alpha_1 | x) < R(\alpha_2 | x)$ than ω_1 else ω_2.

loss loss.

As it is known, Bayes rule minimizes R by

(i) computing $R(\alpha_i | x)$ for every α_i given on x.

(ii) Choosing the action α_i with minimum $R(\alpha_i | x)$

$$R^* = \min R$$

Discriminant function: $g(x) = \ln P(x|\omega_i) + \ln P(\omega_i)$

$\downarrow$ $P(\omega_i|x)$

use ln coz monotonically increasing, does not changes the classification results.

Discriminant function for multivariate Gaussian Density

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

$g_i(x) = \ln P(x|\omega_i) + \ln P(\omega_i)$

now, $g_i(x) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}\ln 2\pi - \frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu) + \ln P(\omega_i)$

Case 1: (Isotropic Gaussian) $\Sigma = \sigma^2 I$

$\Sigma = \sigma^2 I = \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$

Description: In an Isotropic Gaussian, the covariance matrix is scaled identity matrix. This means that the variance of different variable are 'equal' and there is no correlation between variables. Features are statically independent.

We are doing this derivation to find out on which class does a feature(x) belongs to, by finding the ratio of the ln of the posterior of two classes for x.

$$P(x|y=i) = N(\mu_i, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i)\right)$$

For class = 1, taking ln on both side (μ_1)

$$\ln P(x|y=1) = -\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) - \underbrace{\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma|}_{\text{constant} \rightarrow \text{ignore}}$$

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

$$= -\frac{1}{2}\left(\Sigma^{-1} x \cdot x^T - 2x\mu_1^T \Sigma^{-1} + \mu_1^T \mu_1 \Sigma^{-1}\right)$$

$= (x-\mu_i)^T (x-\mu_i)$

$= (x-\mu_i)^2$

$= (x^2 - 2x\mu_i + \mu_i^2)$ $(a^2+2ab+b^2)$

$= (x^T \cdot x - 2x \cdot \mu_i + \mu_i^T \mu_i)\Sigma^{-1}$

Similar for class = 2 just put μ_2

$$\ln P(x|y=2) = -\frac{1}{2}\left(\Sigma^{-1} x \cdot x^T - 2x\mu_2^T \Sigma^{-1} + \mu_2^T \mu_2 \Sigma^{-1}\right)$$

To find the ratio of posterior probability,

$$\ln \frac{P(y=1|x)}{P(y=2|x)} = \ln \frac{P(x|y=1)\,P(y=1)}{P(x|y=2)\,P(y=2)}$$ [evidances are equal & prior are 11]

$$= \ln P(x|y=1) - \ln P(x|y=2)$$

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

$$= -\frac{1}{2}\Sigma^{-1} x \cdot x^T + x\mu_1^T \Sigma^{-1} - \frac{1}{2}\mu_1^T \mu_1 \Sigma^{-1} + \frac{1}{2}\Sigma x \cdot x^T - x\mu_2^T \Sigma^{-1} + \frac{1}{2}\mu_2^T \mu_2 \Sigma^{-1}$$

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

$$= x \quad W^T + W_0$$

$$= W^T x + W_0$$

Thus it converts into linear equation.

$$g_i(x) = g_j(x_0)$$

$$w^T x + w_0 = W^T x_0 + w_0$$

$$\Rightarrow w^T(x - x_0) = 0$$

Case 2: Diagonal Gaussian

$\Sigma_i = \Sigma$.

$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

proof: same as case-1

Case 3: (Full Covariance Gaussian)

Σ_i = arbitrary.

(x) Variance is not equal & has the priors

$$\ln \frac{P(y=1|x)}{P(y=2|x)} = \ln \frac{P(x|y=1)\ P(y=1)}{P(x|y=2)\ P(y=2)}$$

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

$$= -\frac{1}{2}\left[x^T x \Sigma_1^{-1} + x^T x \Sigma_2^{-1} - 2x\mu_1^T \Sigma_1^{-1} + 2x\mu_2^T \Sigma_2^{-1} + \mu_1^T \mu_1 \Sigma_1^{-1} - \mu_2^T \mu_2 \Sigma^{-1}\right] - \ln|\Sigma_1| + \ln|\Sigma_2| + \ln P(y=1) - \ln P(y=2)$$

$$= -\frac{1}{2}\left((\Sigma_1^{-1} - \Sigma_2^{-1}) x^T x\right) - x\left(\mu_1^T \Sigma_1^{-1} - \mu_2^T \Sigma_2^{-2}\right) + \omega_0$$

$$= \omega_{12} x^T x - \omega_{12} x + \omega_0$$

$$= \omega_{12} x^2 - \omega_{12} x + \omega_0$$

= quadratic equation.

So from bayesian we can also derive Quadratic equation

Recursive -40

$$\Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix} \quad \mu_1 = [0\ 0]^T \quad \mu_2 = [3, 3]^T$$

Classify the vector $[1.0\ \ 2.2]^T$

According to mahalonobis distance classifier;

As they share same covariance matrix

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

Avoiding the negative and constant value

$$d_m^2(\mu_1, x) = (x-\mu_1)^T \Sigma^{-1} (x-\mu_1)$$

$$= [1.0\ \ 2.2] \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}^{-1} \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix}$$

$$= [1.0\ \ 2.2] \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix}$$

(1×2) (2×2)

$$= \begin{bmatrix} (1\times 0.95) + (2.2\times(-0.15)) & (-0.15\times 1) + (2.2\times 0.55) \end{bmatrix} \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix}$$

(1×2) (2×1)

$$= [0.62\ \ 1.06] \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix}$$

$$= 2.952$$

$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

$A^{-1} = \frac{1}{ad-cb} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

$\Sigma^{-1} = \frac{1}{1.1\times 1.9 - (0.3)^2} \begin{bmatrix} 1.9 & -0 \\ -0.3 & 1 \end{bmatrix}$

$\Sigma^{-1} = \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix}$

$$dm^2(\mu_2, x) = (x-\mu_2)^T \Sigma^{-1} (x-\mu_2)$$

$$= \begin{bmatrix} -2.0 & -0.8 \end{bmatrix} \begin{bmatrix} 0.05 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} -2.0 \\ -0.8 \end{bmatrix}$$

- i don't know where it come from?

$$= 3.672$$

$$(\mu_2, x) = (x-\mu_2)^T = [1.0 \quad 2.2]^T - [3 \quad 3]^T$$

$$= [-2 \quad -0.8]$$

Zero-One Loss function

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i=j \\ 1 & i \neq j \end{cases}$$

একই হলে সবাই 1 assign করা হবে

otherwise 0

$$R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|x) = \sum_{i \neq j} P(\omega_j|x) = 1 - P(\omega_i|x)$$

So Decision Rule:

Decide ω_1 if $1 - P(\omega_1|x) < 1 - P(\omega_2|x)$; otherwise ω_2

Decide ω_1 if $R(\alpha_1|x) < R(\alpha_2|x)$; " ω_2.

Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$ " ω_2

General loss Vs zero-one Loss

Decide ω_1 if,

$$\frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{(\lambda_{12} - \lambda_{22}) P(\omega_2)}{(\lambda_{21} - \lambda_{11}) P(\omega_1)} \quad \text{Vs} \quad \frac{P(x|\omega_1)}{P(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$$

Case 1: (Isotropic Gaussian)

$$\Sigma = \sigma^2 I = \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

Description: In an Isotropic Gaussian, the covariance matrix is scaled identity matrix. This means that the variance of different variable are 'equal' and there is no correlation between variables. **features are statically independent.

Example: Imagine a circular (sphere) distribution in 2D where the spread of the data is the same along both dimensions.

Properties of Decision boundary:

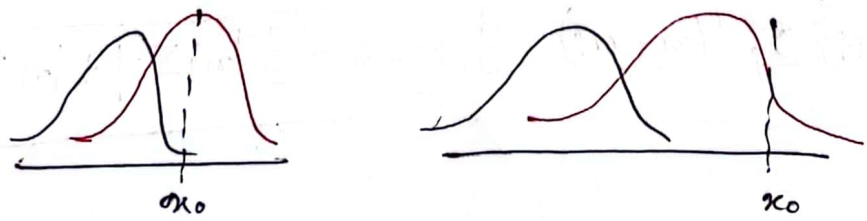
→ It passes through x_0

→ It is orthogonal to the line linking the means

$\mu_1 \perp \mu_2$

→ if $P(\omega_1) \neq P(\omega_2)$ then x_0 shifts away from the most likely category.

→ if σ is very small, the position of the boundary is insensitive to $P(\omega_i)$ and $P(\omega_j)$.



Q: When Bayesian distance become minimum distance classifier.

⇒ Case-1:

When $P(\omega_i)$ are equal then discriminant becomes,

$$g_i(x) = -\frac{\|x-\mu_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

$$g_i(x) = -\|x-\mu_i\|^2$$

[Since all feature has same variance & same Prior Probability it becomes Euclidean D.]

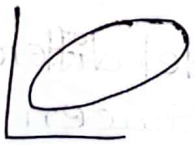
this is the Euclidean distance.

Case 2: (Diagonal Gaussian)

$\Sigma_i = \Sigma$

Description: The covariance matrix is diagonal, indicating that there is no correlation between variables but their variances can be different. This allows for modeling different levels of variability along each dimension.

Example: Think of an Ellipse in 2D where the major and minor axis align with the co-ordinate axes, representing different variances along the two dimensions.



$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

Properties of Decision Boundary:

→ Passes through x_0

→ not orthogonal to the line linking the means.

→ If $P(\omega_i) \neq P(\omega_j)$ then x_0 shifts away from most likely category.

Manhalanobis distance classifier:

→ When $P(\omega_i)$ are equal, then discriminant becomes

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \Sigma^{-1}(x-\mu_i) + \ln P(\omega_i)$$

$$= \frac{-(x-\mu_i)^t(x-\mu_i)}{2\Sigma} + \ln P(\omega_i)$$

$$= \frac{-\|(x-\mu_i)\|^2}{2\sigma^2} + \ln P(\omega_i)$$

$$g_i(x) = -\frac{1}{\sigma^2}\|(x-\mu_i)\|^2 \qquad [\Sigma = \sigma^2]$$

$$= -\frac{\|(x-\mu_i)\|^2}{2\sigma^2} \qquad [\text{since all } P(\omega_i) \text{ are equal so } 1]$$

Here the prior probability does not affect the decision and classification is based solely on the mahalanobis distance term.

Case 3: (Full Covariance Gaussian)

Σ_i = arbitrary.

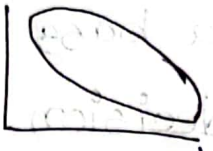
Description: In Full C.G, there is no restrictions on the covariance matrix. It can model different variances and arbitrary correlations between variables. The covariance matrix can have off-diagonal elements, capturing complex relations between variables.

$$\begin{bmatrix} 1 & 0.9 \\ 0.9 & 4 \end{bmatrix}$$

→ clusters have different shapes and sizes (centered at

Example: Envision an ~~eplt~~ ellipse in 2D that is tilted and stretched, indicating correlation or dependence between the two variables.

$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$



D.B is determined by hyperquadrics; setting $g_i(x) = g_j(x)$
(hyperplane, pair of hyperplanes, hyperspheres, hyperellipsoids ...)

Difference:

Case1 Vs Case2: The key difference lies in how they handle correlations. Case1 assumes equal variance and no correlation, while case 2 allows for different variance but still assumes no correlation

Case2 Vs Case3: The case-2 allows different variance but assumes no cross-variable correlations while ~~Both~~ case 3 allows for arbitrary correlations between variables.

Case1 Vs Case3: Case-1 assumes equal variances and no correlation, while ~~Bi~~ case-3 can model different variances and arbitrary correlations, making it more flexible but also requiring more parameters.

Hard margin Vs soft margin

Hard margin:

1. Strict boundary with no tolerance of misclassification.
2. Works well when data is well-seperable but may struggle with noisy or overlapping data.
3. Finds w and b such that,
   $\phi(w) = \frac{1}{2} w^T w$ is minimized and for all $\{(x_i, y_i)\}$
   $y_i (w^T x_i + b) \geq 1$

Soft margin:

1. Allows some misclassification
2. Creates more flexible boundary, can handle noisy or overlapping data.
3. More robust but might sacrifies a bit of accuracy for improved generalization [misclassify ok]
4. $\phi(w) = \frac{1}{2} w^T w + C \sum \xi_i$ is minimized and for all $\{(x_i, y_i)\}$ $y_i (w^T x_i + b) \geq 1 - \xi_i$ and $\xi \geq 0$ for all i

(*) Parameter C control overfitting

A Small C encourage a Wider margin at the expense of allowing some training points to be misclassified, promoting better Generalization.

But a Large C allows the model closely fit the training data which can lead to overfitting and poor generalization to new data. tries to minimize training error so narrow margin and creates overfitting.

Linear SVM Vs Non-Linear SVM

(*) Finds $a_1 \cdots a_N$ such that

$Q(a) = \sum a_i - \frac{1}{2} \sum\sum a_i a_j y_i y_j \boxed{x_i x_j}$

is minimized and

(1) $\sum a_i y_i = 0$

(2) $0 \leq a_i \leq C$ for all a_i

$$\boxed{f(x) = \sum a_i y_i x_i^T x + b}$$

(*) Finds $a_1 \cdots a_N$ such that

$Q(a) = \sum a_i - \frac{1}{2} \sum\sum a_i a_j y_i y_j \boxed{K(x_i, x_j)}$

is minimized and

(1) $\sum a_i y_i = 0$

(2) $a_i \geq 0$ for all a_i

$$\boxed{f(x) = \sum a_i y_i K(x_i, x_j) + b}$$

(***) Higher dimension হোক আর 1D সব জায়গায় dot Product হয়।

Proof: A kernel function is some function that corresponds to an inner product in some expanded feature space.

2-dimension vector: $x = [x_1 \; x_2]$ let $K(x_i, x_j) = (1 + x_i^T x_j)^2$

Need to Prove that $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$; [Linear]

$$K(x_i, x_j) = (1 + x_i^T x_j)^2$$

$$= 1 + x_{i1}^2 x_{j1}^2 + x_{i2}^2 x_{j2}^2 + 2 x_{i1} x_{j1} \cdot x_{i2} x_{j2} + 2 x_{i1} x_{j1} + 2 x_{i2} x_{j2}$$

$$= \begin{bmatrix} 1 & x_{i1}^2 & \sqrt{2} x_{i1} x_{i2} & x_{i2}^2 & \sqrt{2} x_{i1} & \sqrt{2} x_{i2} \end{bmatrix}^T \begin{bmatrix} 1 & x_{j2}^2 & \sqrt{2} x_{j2} x_{j1} & x_{j2}^2 & \sqrt{2} x_{j1} & \sqrt{2} x_{j2} \end{bmatrix}$$

$$= \varphi(x_i)^T \varphi(x_j)$$

SVM

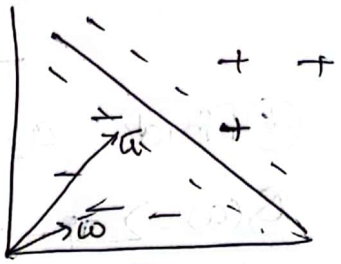


Fig: 1

Proof: for neco unknown point $\vec{u}$ we can write.

$\vec{w} \cdot \vec{u} + b \geq 0$ [from decision Rule]

Now Putting addition constrains to calculate w & b.

$\vec{w} \cdot \vec{x}_+ + b \geq 1 \longrightarrow +ve$

$\vec{w} \cdot \vec{x}_- + b \leq -1 \longrightarrow -ive.$

To combine both condition, neco equation.

$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1$

where, $y_i = +$ for +ve

$y_i = -$ for −ve.

For gutter value x_i,

$$\boxed{y_i(\vec{w}\,\vec{x}_i + b) - 1 = 0}$$

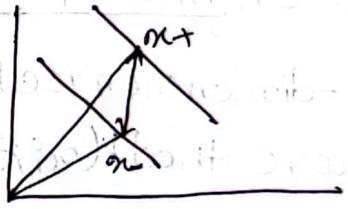


Fig: 2

Again, from Fig: 2

$Wx^+ + b = +1 \Rightarrow Wx^+ + b - 1 = 0 \quad —(i)$

$Wx^- + b = -1 \Rightarrow Wx^- + b + 1 = 0 \quad —(ii)$

From (i) & (ii)

$Wx^+ + b - 1 = Wx^- + b + 1$

$\Rightarrow W(x^+ - x^-) = 2 \quad —(iii)$

From Fig: 2 Margin, $m = x^+ - x^-$ and we need a direct so multiply by a vector, Now m becomes.

$$m = (x^+ - x^-)\frac{W}{\|W\|} \rightarrow 2$$

$$\Rightarrow m = \frac{2}{\|W\|} \quad [\text{From equ (iii)}]$$

Maximize the margins

$$\max \frac{2}{\|w\|} \sim \max \frac{1}{\|w\|} \sim \min \|w\| \sim \min \frac{\|w\|^2}{2}$$

Now minimize margin $\frac{\|w\|^2}{2}$ with respect to $y_i(\bar{\omega}\bar{x}_i + b) - 1$ with Lagrange multipleer.

$$L = \frac{1}{2}\|w\|^2 - \lambda\left[y_i(\bar{\omega}\bar{x}_i + b) - 1\right]$$

$$= \frac{1}{2}\|\omega\|^2 - \Sigma\alpha_i\left[y_i(\bar{\omega}\bar{x}_i + b) - 1\right] \quad \text{— (i)}$$

$$\frac{\partial L}{\partial \omega} = \bar{\omega} - \Sigma \alpha_i y_i x_i = 0 \Rightarrow \bar{\omega} = \Sigma \alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = -\Sigma \alpha_i y_i b = 0 \Rightarrow \Sigma \alpha_i y_i b = 0 \quad \text{— (ii)}$$

Put the values in L equation (1)

$$L = \frac{1}{2}\left(\Sigma \alpha_i x_i y_i\right)\left(\Sigma \alpha_j x_j y_j\right) - \Sigma \alpha_i y_i x_i \left(\Sigma \alpha_j y_j x_j\right) - \Sigma \alpha_i y_i b + \Sigma \alpha_i$$

$$= -\frac{1}{2}\Sigma \alpha_i y_i x_i + \Sigma \alpha_i$$

$$= \Sigma \alpha_i - \Sigma_i \Sigma_j \alpha_i \alpha_j y_i y_j x_i x_j$$

parameters are multiple of each othe so its a quadratic optimization problem.