PAGE :
DATE : / /

16 September 2014

Probability

$P(\text{rain}) = 0.2$ (A)

$P(\text{wet road}) = 0.3$ (B)

$P(A/B) = P(A \cap B)$

if A and B are independent,

$P(A/B) = P(A) \cdot P(B)$

→ conditional probability

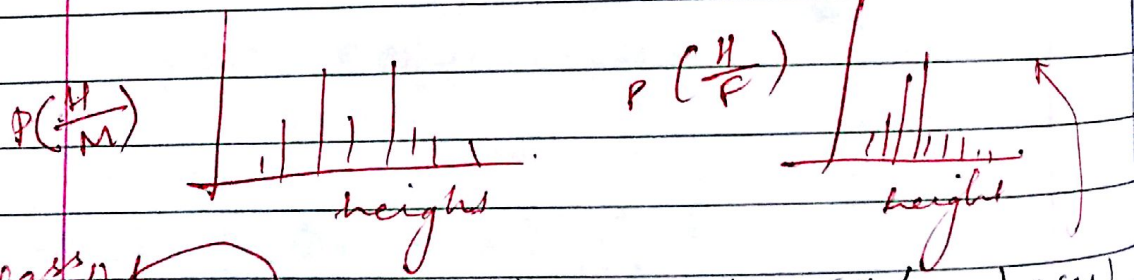
$P(A/B) = \dfrac{P(A \cap B)}{P(B)}$

so $P(A \cap B) = P(A/B) \cdot P(B)$

$= P(B/A) \cdot P(A)$

$P(A/B) = \dfrac{P(B/A)\ P(A)}{P(B)}$

Baye's Theorem



$$P(M/H=64) = \frac{P(H=64/M)\,P(M)}{P(H=64)}$$

$$= \frac{P(H=64/M)\,P(M)}{P(H=64/M)\,P(M) + P(H=64/F)\,P(F)}$$

class label — feature value — likelihood — prior / prejudice — post observation / proposition prob. — evidence — belief

PAGE :
DATE : / /

$$P(F \mid H=64) = 1 - P(M \mid H=64)$$

$$= \frac{P(H=64 \mid F)\,P(F)}{P(H=64 \mid M)\,P(M) + P(H=64 \mid F)\,P(F)}$$

$$P(\omega \mid x) = \frac{P(x \mid \omega)\,p(\omega)}{p(x)}$$

posterior

feature vector

density function

discrete = P

continuous = p

unconditional density of x

19 September 2014

Test for a rare disease (1/million)

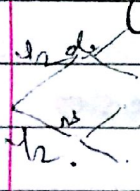
Sensitivity = 100%	False -ve (miss) = 0%	Recall
specificity = 99.9%	False +ve = 0.1%	Precision

→ if you do not have the disease the test will say -ve 99.9% of the time

True +ve

True -ve

if you have a disease you will be caught



$$\frac{0.1}{100} \quad \frac{100}{100+0.1}$$

$$P\left(\frac{+}{D}\right) = 1$$

$$P\left(\frac{+}{ND}\right) = 0.001$$

$$P(D \mid +) = \frac{P(+ \mid D)\,P(D)}{P(+)} =$$

$$= \frac{P(+ \mid D)\,P(D)}{P(+ \mid D)P(D) + P(+ \mid ND)P(ND)} = \frac{10^{-6}}{10^{-6} + 10^{-3}(1-10^{-6})}$$

$$= 10^{-3} = 0.001 = 0.1\%$$

PAGE :
DATE : / /

$P(\text{error}) = ?$

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}, x)\, dx$$

$$= \int_{-\infty}^{\infty} P(\text{error}|x)\, p(x)\, dx$$

if d-dimensional vector, integrate d times of, one over each dimension

Bayes classifier

2 class classifier

$P(\omega_1|X) = 0.7$

$P(\omega_2|X) = 0.3$

we picked ω_1 (> prob)

$P(\text{error}|X) = 0.3$

$P(\text{error}|X) = \min(P(\omega_1|X), P(\omega_2|X))$

Minimize total error, minimize the integral, minimize $P(\text{error}|X)$

Bayes classifier is the best in the world.

→ Nearest Neighbour Classifier

PAGE :
DATE : / /

$$P(\omega_1|X) = \frac{P(X|\omega_1)P(\omega_1)}{P(X)}$$

$$> \frac{P(X|\omega_2)P(\omega_2)}{P(X)}$$

assign it there.

$P(X|\omega_1)P(\omega_1) > P(X|\omega_2)P(\omega_2)$

likelihood prior

Baye's classifier

probability in the world. (impractical)

N parameters, 10N samples.

~~P(error~~ Perror * : Baye's error rate.

23 September 2014

I was Absent!

→ Student statistics.

conditional Groundtruth

Maximum likelihood Estimate

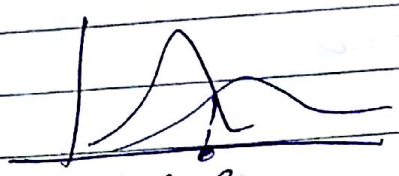
posterior probability $\frac{p(h|w)P(w)}{p(h)} = p(w|h)$

PAGE :
DATE : / /

Gaussian

Now we have prob, we want pdf, how to find it?

lets assume they are Gaussian curves



$\sigma_f = 1.97$ (not much spread)

$\sigma_m = 2.43$ (spread out)

↓

decision point

$x < 66.2$ F

else M

Gaussian: $\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$

for the two classes, μ and σ are different

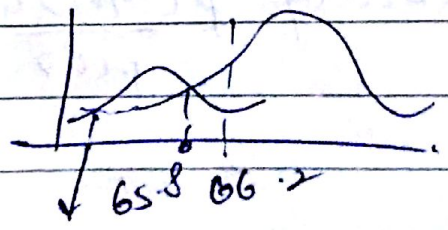
Prior scaled likelihood

$P(x/w)\,P(w)$

case (w_1 by 0.35 → $P(f) = 0.35$
w_2 by 0.65 → $P(m) = 0.65$

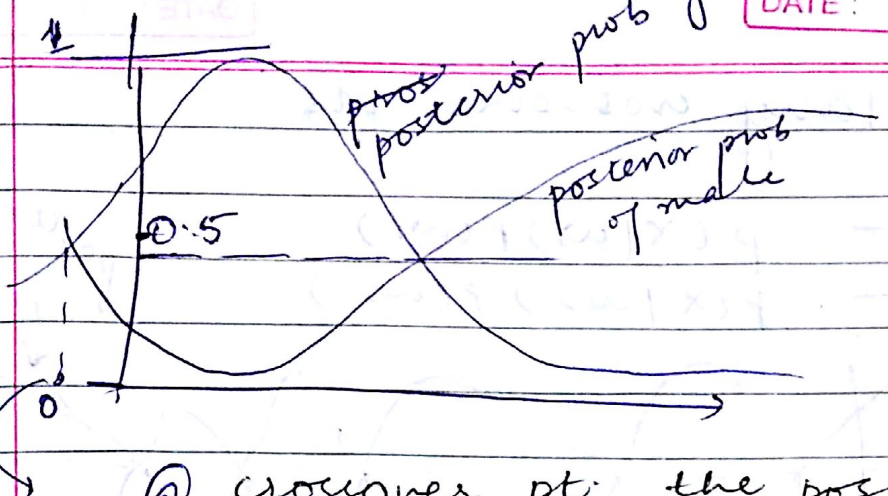
will the pt. of intersection change

Yes



another crossover pt. @ 45.7 : male

$x < 3 \times 10^{11}$

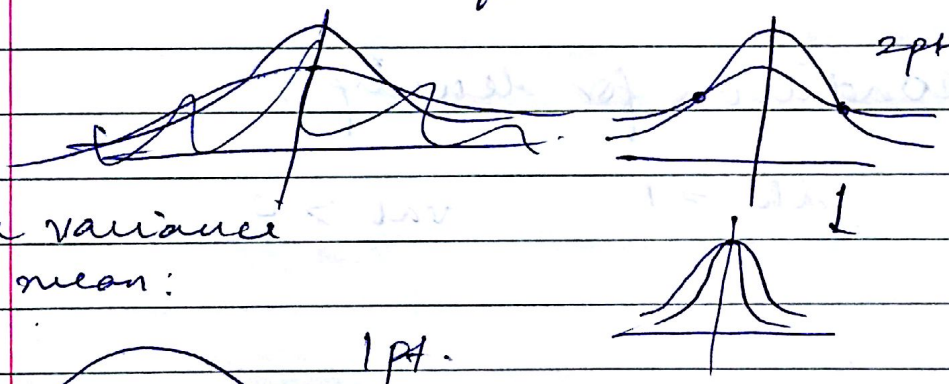


@ crossover pt the posterior prob = 0.5

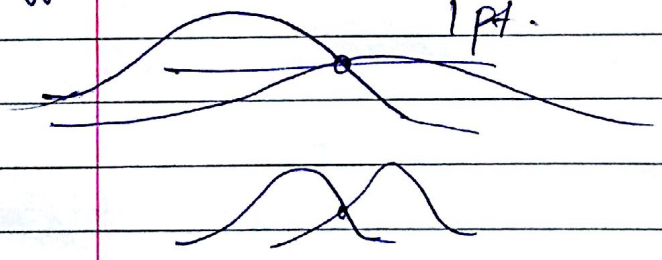
1D - 2 pt of separation : non linear
2D : linear

Bayesian densities can be any densities

Condition for only one pt of separation
same mean, diff variance



same variance
diff mean:

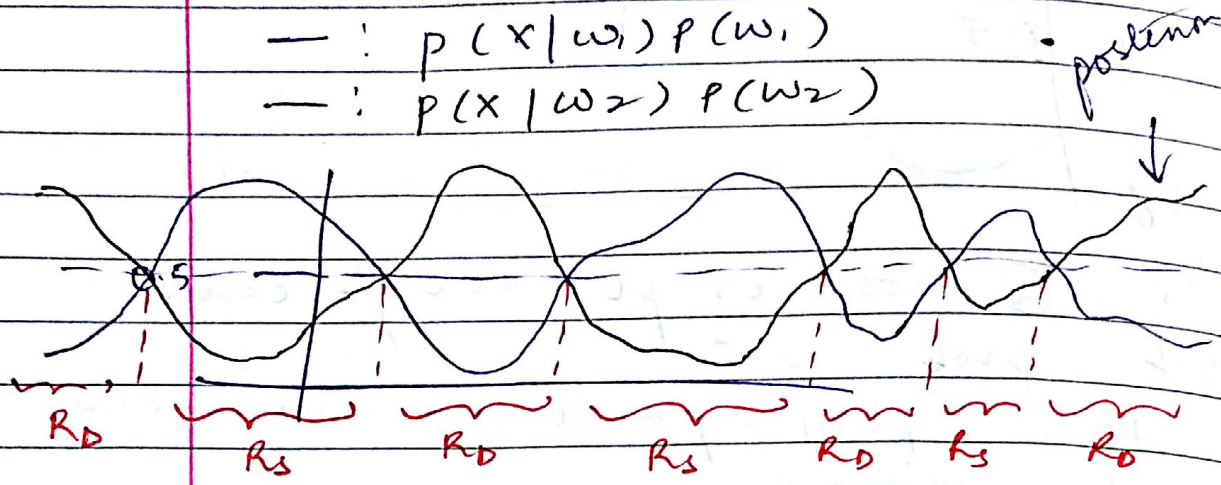


PAGE :
DATE : / /

Many crossover pts.

— : $p(x|\omega_1)P(\omega_1)$

— : $p(x|\omega_2)P(\omega_2)$



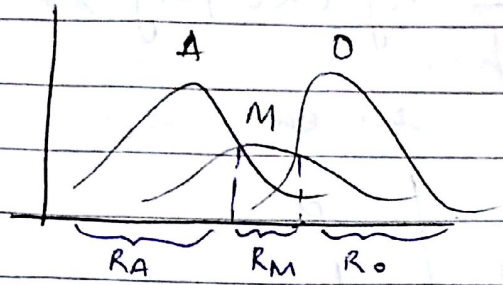
At crossover, value is 0.5 if the curve is going up, it will continue to go up

condition for density:

$\int dh = 1$ val > 0

PAGE :
DATE : / /

26 September 2014



$$P(\omega_1/x) = \frac{P(x|\omega_1)P(\omega_1)}{P(x|\omega_1)P(\omega_1) + P(x|\omega_2)P(\omega_2) + P(x|\omega_3)P(\omega_3)}$$

* Probability theory can be extended to multiple classes.

$$P(\omega_i|x) = \frac{P(x|\omega_i)P(\omega_i)}{\sum_{i=1}^{c} P(x|\omega_i)P(\omega_i)}$$

We are learning each class independently.

→ feature vector : n dimensional
decision boundary : n-1 dimensional
lying in n dimensional space

Equidense surfaces.

→ Giving credit card : loss func.

ω_i	✓	X
CW	0	1
NCW	1	0

(α_i across the top)

credit worthy

$\lambda_{ij}(\alpha_i|\omega_j)$

↑ loss func.

→ zero one loss fun.

minimize loss : $\lambda_{ij} P(\omega_j|x)$

$R(\alpha_i|x)$ ← expected loss

↑ risk (total loss if taking action α_i)

PAGE :
DATE : / /

$$R(\alpha_i|x) = \sum_j \lambda_{ij}(\alpha_i|\omega_j) P(\omega_j|x)$$

action α_i

ground truth ω_j		A	B
	A	0	1
	B	1	0

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

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

take act w that action which minimizes the risk.

α_1 $\lambda_{11} P(\omega_1|x) + \lambda_{12} P(\omega_2|x) < \lambda_{21} P(\omega_1|x) + \lambda_{22} P(\omega_2|x)$

α_1 $(\lambda_{11} - \lambda_{21}) \cdot P(\omega_1|x) < (\lambda_{22} - \lambda_{12}) P(\omega_2|x)$

$$\frac{P(\omega_1|x)}{P(\omega_2|x)} < \frac{\lambda_{22} - \lambda_{12}}{\lambda_{11} - \lambda_{21}}$$

(?) check pw. $\hookrightarrow 1$

likelihood ratio

$$\frac{p(x|\omega_1)}{p(x|\omega_2)} \gtrless \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)}$$

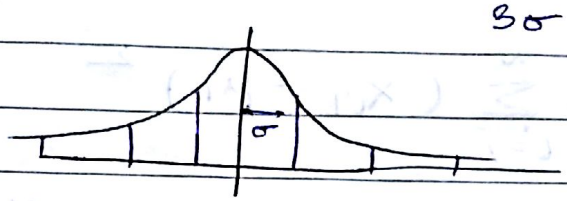
PAGE :
DATE : / /

Gaussian Multivariate Density

$$P(\omega|X) = \frac{p(X|\omega)P(\omega)}{P(X)}$$

$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

3σ : 99.99%



X: vector of two dimension

μ: & vector too!

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{bmatrix}$$

$$\underset{\text{random}}{\underset{\downarrow}{\Sigma}} = \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_d^2 \end{bmatrix}$$

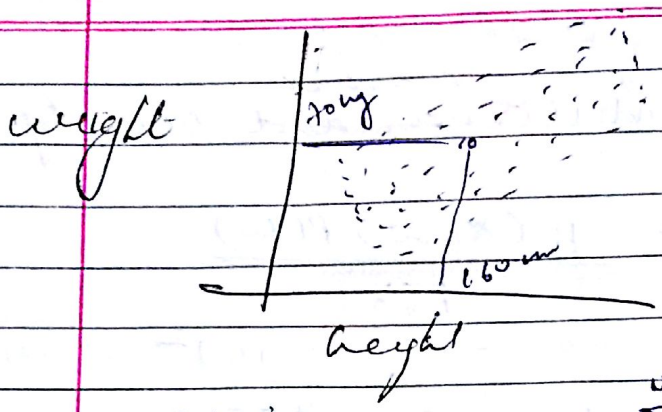
ex Height 160 cm 15 cm σ

70 kg 15 kg

$P(X|\omega)$: (100 cm, 100 kg)

(220 cm, 100 kg)

PAGE :
DATE : / /



variance $= \sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2$

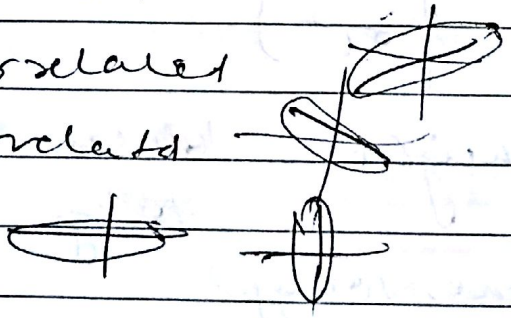
$$\sigma_{11}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_{i1} - \mu_1)^2$$

variance of the quantity x_{i1}
how does the value of x varies around μ.

→ covariance
when 1 quantity change how does the second quant chang.

$$\sigma_{12} = \frac{1}{N} \sum_{i=1}^{N} (x_{i1} - \mu_1)(x_{i2} - \mu_2)$$

+ve correlated
-ve correlated
indep.



$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$ covariance matrix.

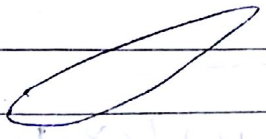
$$[\sigma_{ij}]_{d \times d}$$

PAGE :
DATE : / /

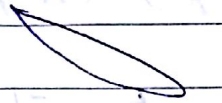
identity covariance matrix :

circle

+ve



-ve



diff diag
off diag 0

27 September 2014

Anoop ??

→ unimodal density

→ Minimum Risk : 2-Category

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \vdots & & & \\ \sigma_{d1} & \cdots & \cdots & \sigma_{dd} \end{bmatrix}_{d \times d}$$

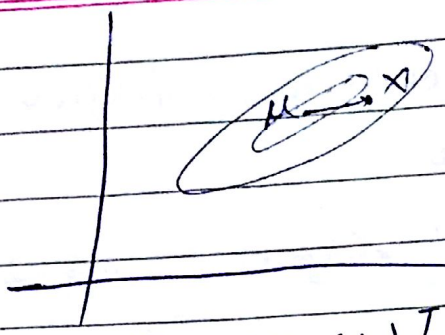
$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix}_{1 \times d}$$

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

↑ univariate

in multivariate case ?

PAGE :
DATE : / /



Euclidian distance $d^2 = (x-\mu)^T(x-\mu)$

$$d_2^2 = (x-\mu)^t \Sigma^{-1} (x-\mu)$$

↳ scaled variant of the distance

P C Mahalanobis

$$N(\mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1} (x-\mu)}$$

Σ : symmetric positive definite

$x^t A x > 0$

dot

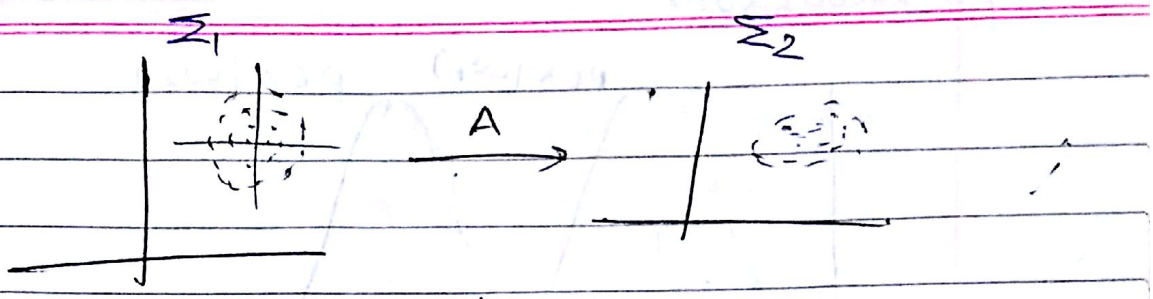
$1 \times d \quad d \times d \quad d \times 1$

1×1

I can do it
Eigen do it

transformed X

PAGE :
DATE : / /



$$\Sigma_2 = A^t \Sigma_1 A$$

if $\Sigma_1 = I$

$$\Sigma_2 = A^t A$$

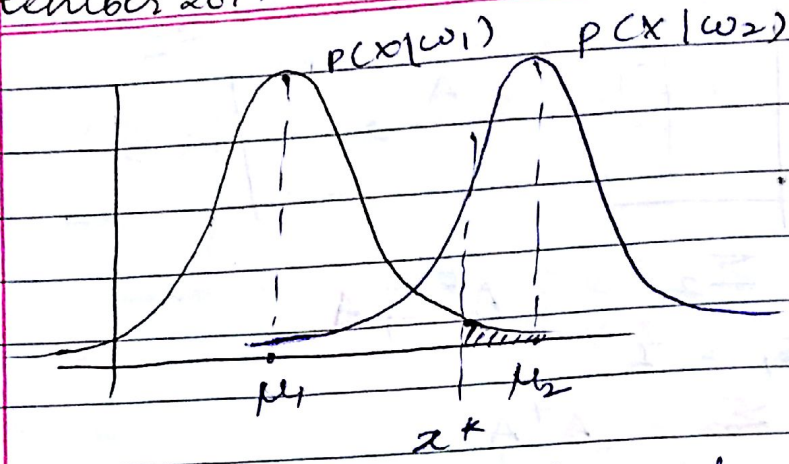
Cholesky Decomposition
for positive definite symmetric matrix hermitian.

Book se

2.6 Discriminant func. for one Normal density
log / sqrt / any monotonic func.
for normal dist.

PAGE :
DATE : / /

30 September 2014



Because of a noisy channel, we send μ_1 but we get μ_1 + noise.

x^* : error / noise

what should unlock cannot $1 \to 0$
what shouldn't unlock can $0 \to 1$

x^* : noisy decision boundary

① $x > x^*$, ω_2 : hit
② $x < x^*$, ω_2 : miss
③ $x > x^*$, ω_1 : false alarm
④ $x < x^*$, ω_1 : correct rejection

$0 \to 0$ (ω_2 41)

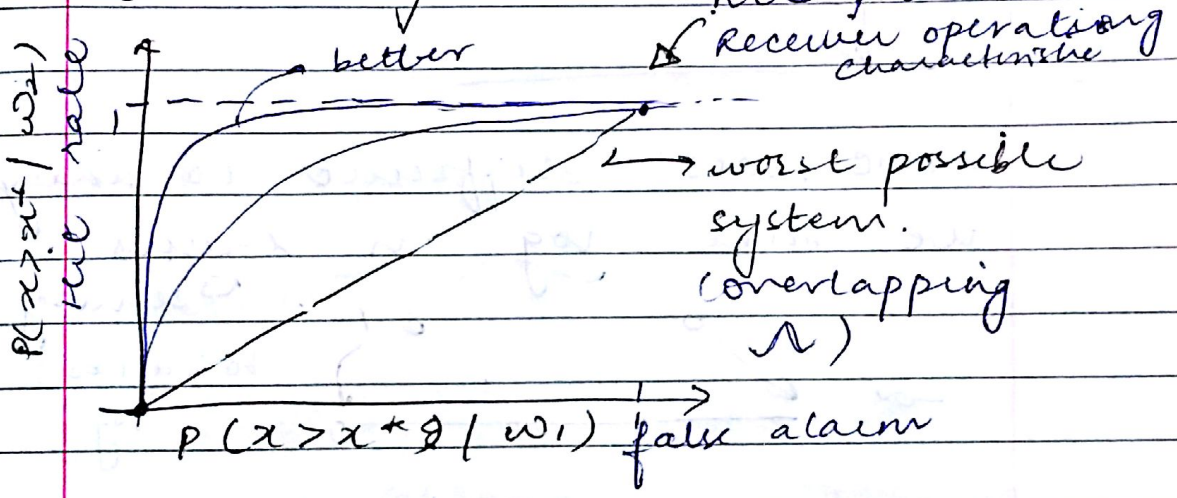
	0 0 →	correct rejection
	0 1 →	false alarm
	1 0 →	miss
	1 1 →	hit

What is the prob. of a false alarm, given threshold is x^*

PAGE :
DATE : / /

prob. when density is given : area under the curve.

To understand the classifier, we need one of hit; miss and one of false alarm and correct rejection. ROC plot



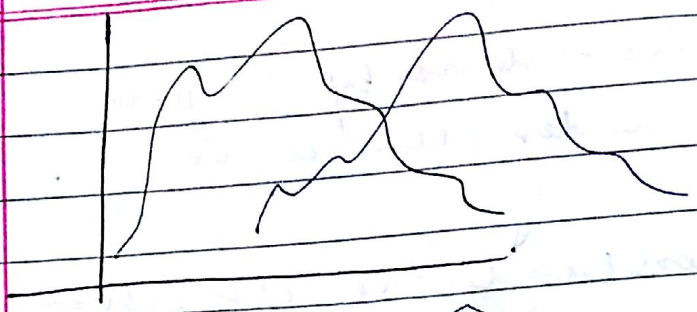
operating point of the system dependent on the scenario.

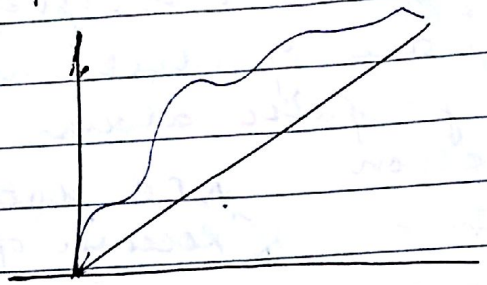
$$d' = \frac{|\mu_2 - \mu_1|}{\sigma} \rightarrow \text{or } \frac{\sigma_1 + \sigma_2}{2}$$

more the d' : better the curve

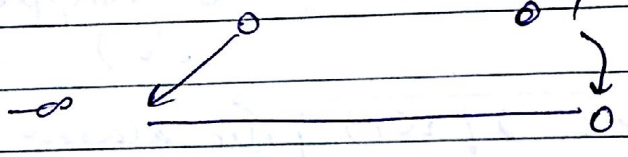
The same idea works even if the distribution is not normal

PAGE :
DATE : / /

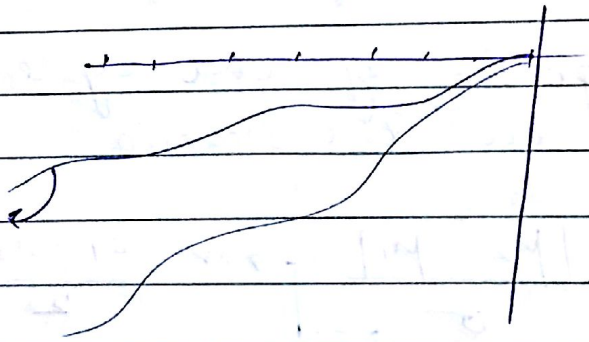




2 ROC plots: difficult to analyze
we take log in x-axis
↳ semilog



both axes: log.



Equal Error Rate

Confusion Matrix

o/p \ Ground truth	ω_1	ω_2	ω_3
ω_1			
ω_2			
ω_3			

best : diag = 1 off diag = 0

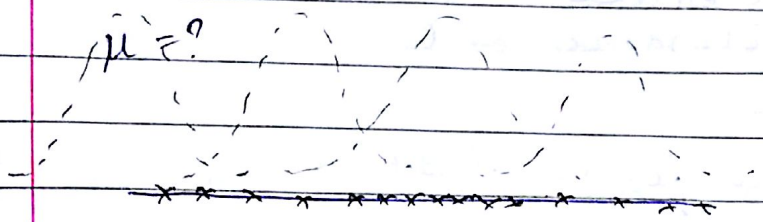
PAGE :
DATE : / /

Practically impossible to learn a complex density function.

we know that dist is normal,

$\sigma = 10$.

$\mu = ?$



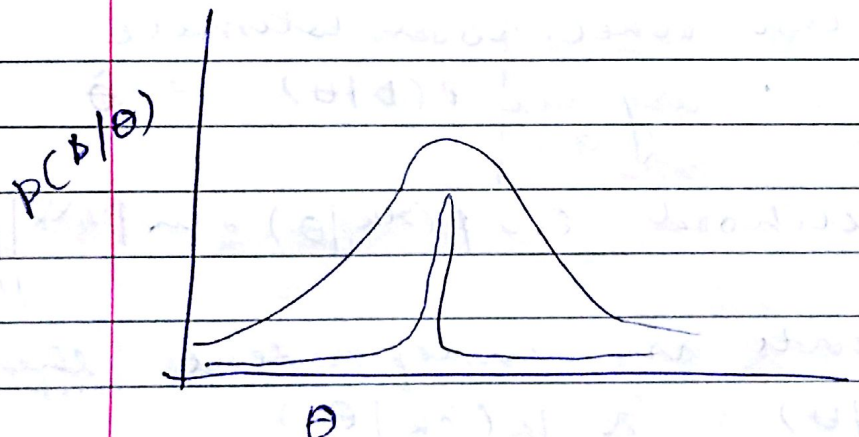
$$P(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

D → complete dataset

θ → μ

sampling is iid

independent & identically distributed



likelihood of theta generating the samples.

PAGE :
DATE : / /

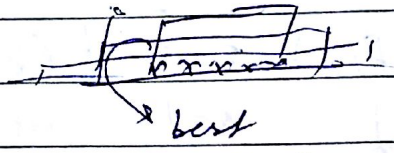
$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p(D|\theta)$

$\hat{\theta}$ ↓ estimated

$\hat{\theta}_{MLE}$ (MLE)

Maximum estimate likelihood dissemination

for uniform dist



$a = \min\{D\}$

$b = \max\{D\}$

7 October 2014

Maximum likelihood Estimate

(MLE) : $\underset{\theta}{\operatorname{argmax}}\ P(D|\theta) = \hat{\theta}$

log likelihood $\ln p(x_k|\theta) = \ln p(x_k|\mu,\Sigma)$ μ,Σ

if events are independent then,

$P(D|\theta) = \prod_{i=1}^{n} p(x_k|\theta)$

PAGE :
DATE : / /

$$p(x_k|\mu,\Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{\frac{1}{2}(x_k-\mu)^t \Sigma^{-1}(x_k-\mu)}$$

$$\ln p(x_k|\mu,\Sigma) = -\frac{1}{2}\ln[(2\pi)^d|\Sigma|] - \frac{1}{2}(x_k-\mu)^t\Sigma^{-1}(x_k-\mu)$$

For MLE :
differentiate and put to '0'

$$\nabla_\mu(\ln(p(x_k|\mu))) = -\frac{1}{2} \times \Sigma^{-1} \times -2(x_k-\mu)$$

$$\begin{bmatrix}(x_1-\mu_1)\\(x_2-\mu_2)\\\vdots\\(x_d-\mu_d)\end{bmatrix}^t \begin{bmatrix}(x_1-\mu_1)\\\vdots\\(x_d-\mu_d)\end{bmatrix} = \begin{matrix}(x_1-\mu_1)^2+\\(x_2-\mu_2)^2+\\(x_3-\mu_3)^2+\cdots\\(x_d-\mu_d)^2\end{matrix}$$

$$\nabla_\mu(\cdot) = \begin{bmatrix}-2(x_1-\mu_1)\\-2(x_2-\mu_2)\\\vdots\\-2(x_d-\mu_d)\end{bmatrix} = -2(X-\mu)$$

$$= \Sigma^{-1}(X_k-\mu)$$

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

$$\therefore \nabla_\mu(\ln p(D|\theta)) = \sum_{k=1}^{n} \Sigma^{-1}(x_k-\hat{\mu}) = 0$$

summation — covariance matrix

PAGE :
DATE : / /

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

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

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

case 2: μ and σ are unknown [1-D]

$\theta_1 = \mu \qquad \theta_2 = \sigma^2$

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

$$\ln p(x_k|\theta) = \frac{1}{2} \ln 2\pi|\theta_2|$$

$$= -\frac{1}{2} \ln 2\pi\theta_2 - \frac{1}{2} \frac{(x_k - \mu)^2}{\sigma^2}$$

$$= -\frac{1}{2} \ln 2\pi\theta_2 - \frac{1}{2} \frac{(x_k - \theta_1)^2}{\theta_2}$$

$$\nabla_\theta \ln(p(x_k|\theta)) = \begin{bmatrix} \frac{x_k - \theta_1}{\theta_2} & (1) \\ \frac{-1}{2\theta_2} + \frac{1}{2} \frac{(x_k - \theta_1)^2}{\theta_2^2} & (2) \end{bmatrix}$$

$$\frac{-1}{2} \ln 2\pi - \frac{1}{2} \ln \theta_2$$

$$-\frac{1}{2} \times \frac{1}{\theta_2} \qquad \frac{d}{dx} x^{-1} = -1 x^{-2}$$

PAGE :
DATE : / /

at $\hat{\theta}$

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

$\hat{\theta}_2 \neq 0$

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

and

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

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

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

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

Multivariate MLE for $N(\mu, \Sigma)$

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

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})(x_k - \hat{\mu})^t$$

unbiased ↳ $E\left[\frac{1}{n}\sum_{k=1}^{n} x_k\right] = \mu$

$E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \neq \sigma^2$

$= \left(\frac{n-1}{n}\right)\sigma^2$

↓ biased estimate

PAGE :
DATE : / /

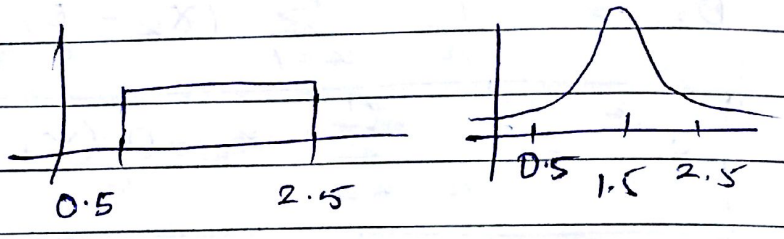
$$E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \sigma^2$$

unbiased (not MLE)

10 October 2014

1. $p(x|\theta)$ is known, θ is unknown — MLE / MAP

2. $p(\theta)$: all prior knowledge about θ
(all prior knowledge about θ is encoded into a density func.)



$$P(\omega_i|x,D) = \frac{p(x|\omega_i,D)\,P(\omega_i|D)}{\sum_{i=1}(\ldots)}$$

$p(x|D_i)$

$p(x|D)$ → learn this part

$$p(x|D) = \int p(x,\theta|D)\,d\theta$$

$$= \int p(x|\theta,D)\,p(\theta|D)\,d\theta$$

$$= \int \underbrace{p(x|\theta)}_{\text{known}}\,p(\theta|D)\,d\theta$$

PAGE :
DATE : / /

$$p(x|D) = \int p(x|\theta)\, p(\theta|D)\, d\theta$$

for MLE :

$p(\theta|D)$ is zero at all θ except $\hat{\theta}$

so,

$$p(x|D) = p(x|\hat{\theta})$$

D all samples $\in \omega_i$

$= p(x|\omega_i)$

do it for each class

→ ← prior

$$p(\theta|D) = \frac{p(D|\theta)\, p(\theta)}{p(D)}$$

$$= \frac{p(D|\theta)\, p(\theta)}{\frac{1}{\alpha} = \int p(\theta, D|\theta)\, p(\theta)\, d\theta}$$

case 1 $\theta = \mu$ σ = known

prior of μ is modelled as

$p(\mu) \sim N(\mu_0, \sigma_0^2)$

$$\propto p(D|\mu)\, p(\mu) = \alpha \left[\prod_{k=1}^{n} p(x_k|\mu)\right] p(\mu)$$

indep & identically dist

$$= \alpha \left[\prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2}\right] \cdot \frac{1}{\sqrt{2\pi}\,\sigma_0} e^{-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2}$$

PAGE :
DATE : / /

$$-\frac{1}{2}\left[\sum_{k=1}^{n}\left(\frac{x_k-\mu}{\sigma}\right)^2+\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]$$

α' e

$$-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{\mu_0}{\sigma_0^2}\right)\mu\right]$$

α'' e

$$-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{\mu_0}{\sigma_0^2}+\frac{1}{\sigma^2}\sum_{k=1}^{n}x_k\right)\mu\right]$$

α'' e

$$= p(\mu/D) \sim N(\mu_n, \sigma_n)$$

when prior was normal density, $p(\mu/D)$ is also a normal density. The shape of our belief remains the same.

$$\mu_n = \frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}$$

(3) $$p(x/D) = \int p(x/\theta)\, p(\theta/D)\, d\theta$$

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

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2} = \frac{1}{\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)}$$

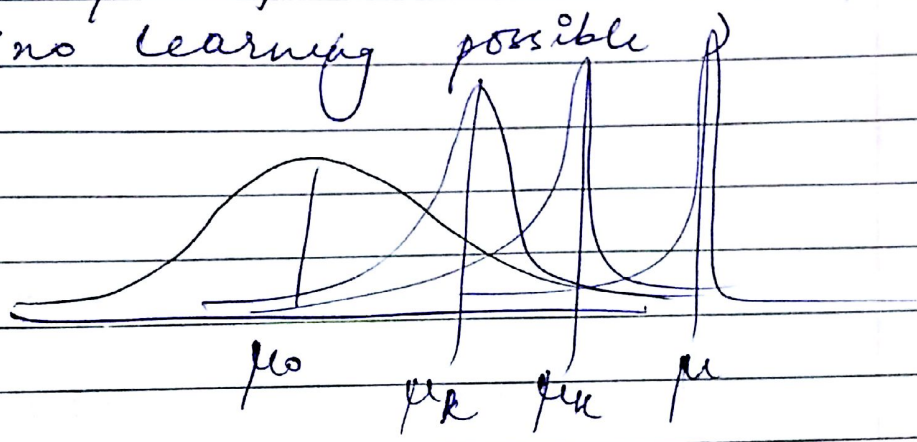
$\bar{x}_n$ → mean of the samples

PAGE :
DATE : / /

n = large (prior not imp.)

σ^2 = large (unable to learn)

$\sigma_0 = 0$: fundamentalist
(believe in something that ~~and~~ is completely certain)
(no learning possible)



~~Bayes~~ Bayesian Learning

prior - Normal : Conjugate Prior
posterior - Normal
reproducing density