Lectures 3 & 4: April 4 and 6

- Lectures 3 and 4: Covered: 3.1-3.4, 4.1-4.2
- Probabilistic view: Classification & Hypotheiss Testing
  - Uncertainy in evidence and predictors
- Maximum Likelihood, LRT
- Bayes Estimation
  - Maximum A Posterior
  - Minimum Expected Bayes Risk
- ROC curves
- Example : Pfa and Pmd

- iIssues with Bayes Decision Theory
  - Why and when?
  - Empirical Bayes Risk

Bayesian Detection

→ Review from prev. lecture

→ y = unknown class label, $y = 0$ or 1

→ x = data / evidence

→ Problem: Estimate y from x

→ Denote $\hat{y}$ = estimate of y.

→ Probabilistic set up:

Assume we know "likelihoods"

$p(x|y=0)$ and $p(x|y=1)$

→ Distribution of evidence given y.

ML Estimate

$$\hat{y} = \arg\max_{y} p(x|y)$$

→ Selects y that ~~leads to~~ for which x is most likely.

→ Binary case

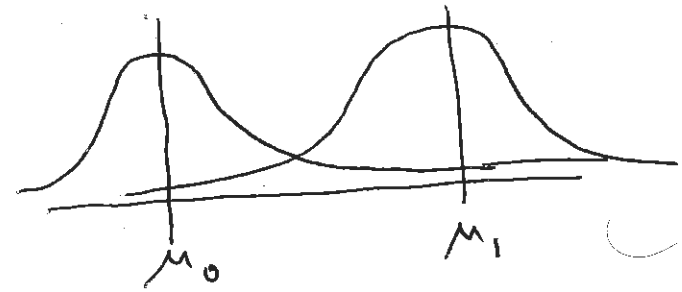
$$\hat{y} = \begin{cases} 1 & \text{when } p(x|y=1) \geq p(x|y=0) \\ 0 & \text{when } p(x|y=1) \leq p(x|y=0). \end{cases}$$

Example ML classifier for two Gaussians

$y = 0$ or 1

$y=0 \Rightarrow x \sim \mathcal{N}(\mu_0, \sigma^2)$
$y=1 \Rightarrow x \sim \mathcal{N}(\mu_1, \sigma^2)$

Assume $\mu_1 > \mu_0$. Note σ^2 same



MLE classifier

$$p(x|y=i) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu_i)^2/2\sigma^2}$$

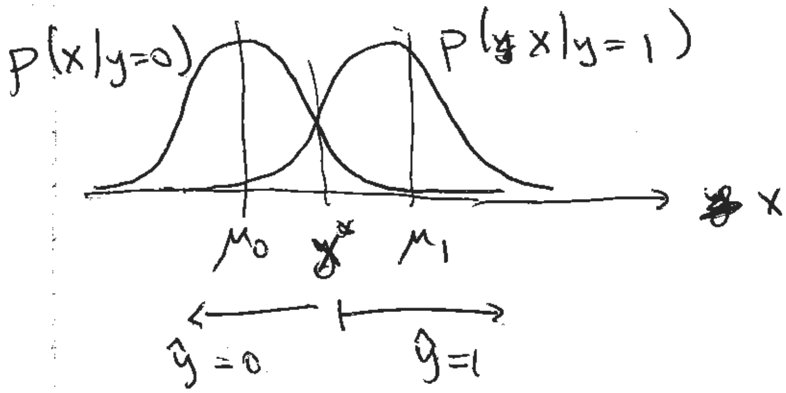
$$\hat{y}=1 \iff p(x|y=1) \geq p(x|y=0)$$

$$\iff \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(y-\mu_1)^2/2\sigma^2} \geq \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(y-\mu_0)^2/2\sigma^2}$$

$$\iff (y-\mu_1)^2 \leq (y-\mu_0)^2$$

$$\iff y^2 - 2\mu_1 y + \mu_1^2 \leq y^2 - 2\mu_0 y + \mu_0^2$$

$$\iff y \geq \frac{\mu_1^2 - \mu_0^2}{2(\mu_1 - \mu_0)} = \frac{\mu_1 + \mu_0}{2} = x^* = \text{midpoint}$$



$$\hat{y} = \begin{cases} 1 & \text{if } x \geq x^* \\ 0 & \text{if } x \leq x^* \end{cases}$$

How to decide Sea Bass or Salmon?

Airplane or Bird

Maximum Likelihood (ML)

$\hat{y}_{ML} = \underset{y}{\text{ARG MAX}}\ p(x|y)$ $\quad \left(p(x|\hat{y}_{ML}) \geq p(x|y) \ \forall y \right)$

If $p(x|y=1) > p(x|y=-1)$ decide $y=1$

otherwise $y=-1$

Equivalently $\log \frac{p(x|y=1)}{p(x|y=-1)} > 0$ log-likelihood test.

Seems reasonable, but what if birds are more likely than airplanes?

Must take into account the prior probability $p(y=1), p(y=-1)$.

Bayes Rule $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$

prob of y conditioned on observation.

If $p(y=1|x) > p(y=-1|x)$ decide $y=1$

otherwise decide $y=-1$

Maximum a Posteriori (MAP) $\hat{y}_{MAP} = \underset{y}{\text{ARG MAX}}\ p(y|x)$

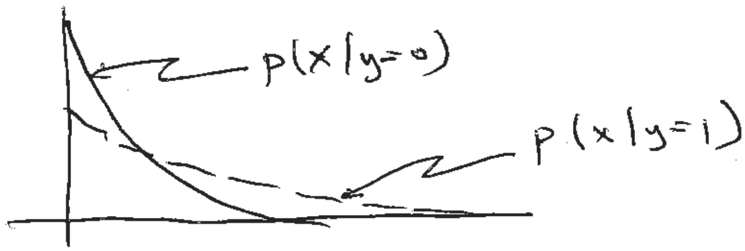
③

Example $\quad y = 0$ or 1

→ $p(x|y=i) = \frac{1}{\lambda_i} e^{-x/\lambda_i}$, $\lambda_i = E(x|y=i)$

→ Two exponentials

→ Assume $\lambda_1 \geq \lambda_0$



→ Do LRT in log domain

$$\cancel{\ln P}\quad L(x) = \ln\left[\frac{p(x|y=1)}{p(x|y=0)}\right]$$

$$= x\left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right) + \ln\left(\frac{\lambda_0}{\lambda_1}\right)$$

LRT

$$L(x) \geq \gamma \iff x\left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right) + \ln\left(\frac{\lambda_0}{\lambda_1}\right) \geq \gamma$$

$$\iff x \geq x^* = \underbrace{\left(\frac{1}{\lambda_0} - \frac{1}{\lambda_1}\right)^{-1}\left[\gamma + \ln\left(\frac{\lambda_1}{\lambda_0}\right)\right]}_{\text{Threshold.}}$$

→ Linear classifier.

(5) Bayes Risk

Bayes Risk is the best you can do if: (a) you know $p(x|y)p(y)$ & $L(\cdot,\cdot)$

(b) you can compute $\hat{\alpha} = \arg\min_{\alpha} R(\alpha)$.

(c) you can afford the losses (e.g. gambling, poker)

(d) you make the decision for a sequence of data $x_1 \ldots x_n$ with states $y_1 \ldots y_n$ where each (x_i, y_i) are independently identically distributed from $p(x,y)$

Bad – if you are playing a game against an intelligent opponent. (use Game Theory instead)

Bad – if any of (a), (b), (c), (d) are wrong

Note: Cognitive Scientists study whether people use decision theory. Kahneman & Tversky argue that people do not – Prospect Theory. Debatable.

(4) Risk

The risk of the decision rule $\alpha(x)$ is the expected loss.

$$R(\alpha) = \sum_{x,y} L(\alpha(x), y)\, P(x,y)$$

(Note integrate $\int dx$ if x is contenuies)

Bayes Decision Theory says "pick the decision rule $\hat{\alpha}(\cdot)$ which minimizes the risk".

$$\hat{\alpha} = \underset{\alpha \in A}{\mathrm{ARGMIN}}\, R(\alpha), \quad R(\hat{\alpha}) \geq R(\alpha) \ \forall\, \alpha \in A.$$

A = set of all decision rules

$\hat{\alpha}$ is Bayes Decision

$R(\hat{\alpha})$ is Bayes Risk.

I Review of Bayes' Risk

$y = 0$ or 1 = unknown class label

x = data or evidence

→ $P(y)$ = prior

→ $P(x|y)$ = likelihood

→ $L(\hat{y}, y)$ = loss fn.

$\hat{y} = \alpha(x)$ = estimator

$$R(\alpha) = \sum_{x,y} L(\alpha(x), y) \underbrace{P(x,y)}_{=P(x|y)P(y)} \quad \leftarrow \text{Bayes' Risk}$$

$$\hat{\alpha} = \arg\min_{\alpha} R(\alpha) \leftarrow \text{Min. Bayes' risk est.}$$

$$\hat{\alpha}(x) = \arg\min_{\hat{y}} \sum_{y} L(\hat{y}, y) \underbrace{P(y|x)}_{\text{posterior}} = \arg\min_{\hat{y}} E(L(\hat{y},y)|x)$$

Binary case $\quad y = 0$ or 1, $\hat{y} = 0$ or 1 $\quad L(\hat{y} = i, y = j) = C_{ij}$

$\hat{\alpha}(x) = 1$

$$\Leftrightarrow C_{11} P(y=1|x) + C_{10} P(y=0|x) \le C_{01} P(y=1|x) + C_{00} P(y=0|x)$$

Special case $C_{11} = C_{00} = 0$.

$$\hat{\alpha}(x) = 1 \Leftrightarrow \frac{C_{01} P(y=1|x)}{C_{10} P(y=0|x)} \ge 1$$

Bayes' rule $\quad P(y|x) = \dfrac{P(x|y)\,P(y)}{P(x)}$

Hence

$$\hat{\alpha}(x) = 1 \iff \underbrace{\frac{P(x|y=1)}{P(x|y=0)}}_{L(x) = \text{Likelihood ratio}} \geq \underbrace{\frac{P(y=0)}{P(y=1)}\,\frac{C_{10}}{C_{01}}}_{t = \text{threshold}}$$

Likelihood Ratio Test

$$L(x) = \frac{P(x|y=1)}{P(x|y=0)}$$

$$\hat{\alpha}(x) = \begin{cases} 1 & L(x) \geq t \\ 0 & L(x) < t \end{cases}$$

→ Any Bayes' opt. test is an LRT for some t.

~~→ Can compute~~

→ MAP & ML also are LRT's.

ML $\quad L(x) \gtrless 1$

MAP $\quad L(x) \gtrless \dfrac{P(y=0)}{P(y=1)}$

6) Better understanding of Bayes Decision Theory. Re-express

$$R(\alpha) = \sum_x \sum_y L(\alpha(x), y) P(x,y)$$

$$= \sum_x P(x) \left\{ \sum_y L(\alpha(x), y) P(y|x) \right\}$$

Hence, for each x,

$$\hat{\alpha}(x) = \underset{\alpha(x)}{\text{ARG MIN}} \sum_y L(\alpha(x), y) P(y|x)$$

Obtaining MAP & ML as special cases.

If $y \in \{-1, 1\}$ and the loss function penalizes all errors equally:

$$L(\alpha(x), y) = 1, \text{ if } \alpha(x) \neq y.$$
$$= 0, \text{ otherwise}$$

Then $\hat{\alpha}(x) = \underset{\alpha(x)}{\text{ARG MAX}}\ P(y = \alpha(x) | x)$ $y \in \{-1, 1\}$

MAP estimate.

If also $p(y=1) = p(y=-1)$, then

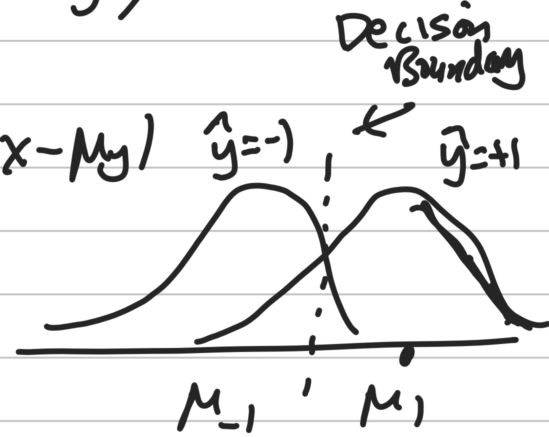
$\hat{\alpha}(x) = \underset{\alpha(x)}{\text{ARG MAX}}\ p(x | y = \alpha(x))$ ML estimate

(7) <u>Examples</u> $p(x|y) = \frac{1}{\sqrt{2\pi}\,\sigma_y} e^{-\frac{(x-\mu_y)^2}{2\sigma^2}}$.

$y \in \{-1, 1\}$ $p(y) = \frac{1}{2}$.

$L(\alpha(x), y) = 1$, if $\alpha(x) \neq y$, $= 0$ otherwise.

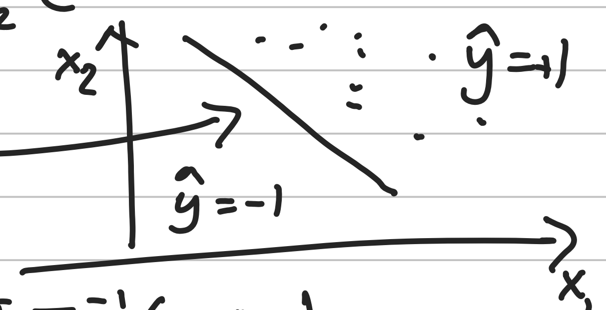
Bayes Rule $\alpha(x) = \underset{y \in \{-1,1\}}{\text{ARG MIN}} |x - \mu_y|$



<u>Suppose</u> $\underline{x}$ is a vector in two dimensions

$$p(\underline{x}|y) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}|\underline{x} - \underline{\mu}_y|^2}$$

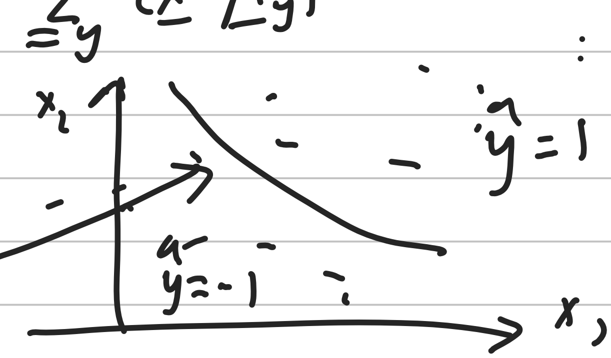
separting line/plane → Decision Boundary



if $p(\underline{x}|y) = \frac{1}{2\pi|\underline{\Sigma}_y|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_y)^T \underline{\Sigma}_y^{-1}(\underline{x}-\underline{\mu}_y)}$

Gaussians with unequal covariance.

Decision Boundary



More Details

$$P(\underline{x}|y) = \frac{1}{2\pi|\underline{\underline{\Sigma}}|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_y)^T \underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu}_y)}$$

ie same covariance $\underline{\underline{\Sigma}}$ for both classes $y=\pm 1$.

$$\log \frac{P(\underline{x}|y=1)}{P(\underline{x}|y=-1)} = \frac{1}{2}(\underline{x}-\underline{\mu}_{-1})^T \underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu}_{-1}) - \frac{1}{2}(\underline{x}-\underline{\mu}_1)^T \underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu}_1)$$

($2\pi|\Sigma|^{1/2}$ terms cancel)

Linear in $\underline{x}$ desribes a plane.

$$= (\underline{\mu}_{-1}-\underline{\mu}_1)^T \underline{\underline{\Sigma}}^{-1}\underline{x} + \frac{1}{2}\underline{\mu}_{-1}^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_{-1} - \frac{1}{2}\underline{\mu}_1^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 .$$

Hence ML rule/estimator corresponds to a rule.

Classify $\underline{x}$ as $y=1$ if

$$(\underline{\mu}_{-1}-\underline{\mu}_1)^T \underline{\underline{\Sigma}}^{-1}\underline{x} + \frac{1}{2}\underline{\mu}_{-1}^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_{-1} - \frac{1}{2}\underline{\mu}_1^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 > 0$$

as $y=-1$ if $(\underline{\mu}_{-1}-\underline{\mu}_1)^T \underline{\underline{\Sigma}}^{-1}\underline{x} + \frac{1}{2}\underline{\mu}_{-1}^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_{-1} - \frac{1}{2}\underline{\mu}_1^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 < 0$.

If there is a prior $p(y)$

$$\log \frac{P(y=1|\underline{x})}{P(y=-1|\underline{x})} = \log \frac{P(\underline{x}|y=1)P(y=1)}{P(\underline{x}|y=-1)P(y=-1)}$$

$$= \log \frac{P(\underline{x}|y=1)}{P(\underline{x}|y=-1)} + \underbrace{\log \frac{P(y=1)}{P(y=-1)}}$$ $\leftarrow$ Indep of $\underline{x}$

($P(y|\underline{x}) = \frac{P(\underline{x}|y)P(y)}{P(\underline{x})}$, $p(\underline{x})$ cancels in the ratio.)

Hence prior shifts the seperating plane to

$$(\underline{\mu}_{-1}-\underline{\mu}_1)^T \underline{\underline{\Sigma}}^{-1}\underline{x} + \frac{1}{2}\underline{\mu}_{-1}^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_{-1} - \frac{1}{2}\underline{\mu}_1^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 + \log\frac{P(y=1)}{P(y=-1)} .$$

With Loss Function

$$R(\alpha(\underline{x})=1) = L(1,1)P(y=1|\underline{x}) + L(1,-1)P(y=-1|\underline{x})$$

$$R(\alpha(\underline{x})=-1) = L(-1,1)P(y=1|\underline{x}) + L(-1,-1)P(y=-1|\underline{x})$$

Decision boundary occurs where $R(\alpha(\underline{x})=1) = R(\alpha(\underline{x})=-1)$.

ie. when $\{L(1,1)-L(-1,1)\}P(y=1|\underline{x}) = \{L(-1,1)-L(1,-1)\}P(y=-1|\underline{x})$

ie when

$$\log\frac{P(y=1|\underline{x})}{P(y=-1|\underline{x})} = \log\frac{\{L(-1,1)-L(1,-1)\}}{\{L(1,1)-L(-1,1)\}}$$

$\nearrow$ additional shift in position of seperating plane.

(4) ROC curves

Note Title

Bayes Decison Theory. $R(\alpha) = \sum_{x,y} L(\alpha(x), y) P(x,y)$

For binary $y \in \{\pm 1\}$

The Bayes Rule $\hat{\alpha} = \arg\min_{\alpha} R(\alpha)$

reduces to thresholding the log-likelihood ratio. i.e. it is of form:

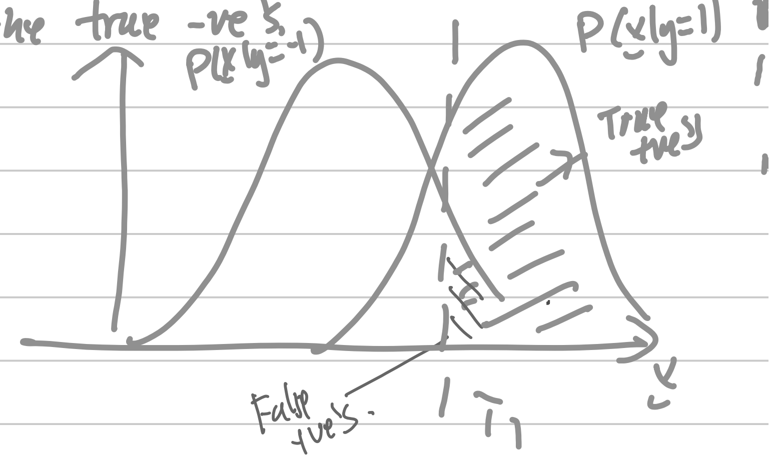
$$\hat{\alpha}_T(x) = 1, \quad \text{if} \quad \log \frac{P(y=1|\underline{x})}{P(y=-1|\underline{x})} > T$$

$$\hat{\alpha}_T(x) = -1, \quad \text{otherwise}$$

The threshold T is a function of the prior $P(y)$ and the loss function $L(\alpha(x), y)$. Hence changing the prior, or the loss function, corresponds to changing T.

So $\log \frac{P(y=1|\underline{x})}{P(y=-1|\underline{x})}$ is an example of the function $f(\underline{x})$ (previous page)

Changing T will alter the false +ve's, the true +ve's, the false -ve's, the true -ve's.

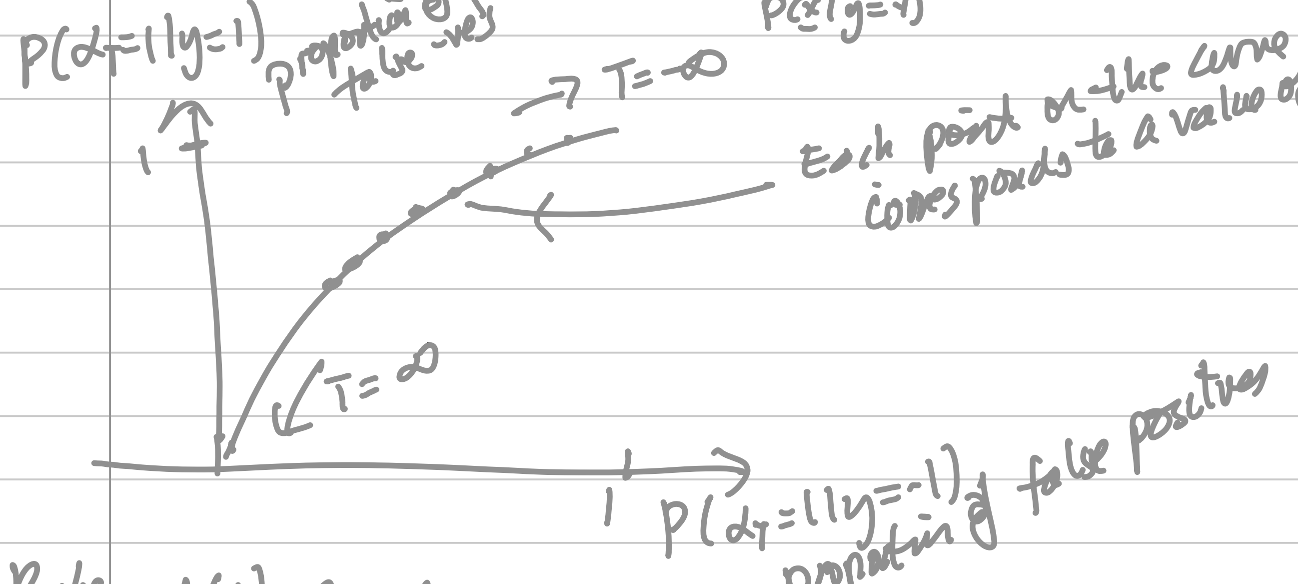


(5)

ROC curve. Plot $P(\alpha_T = 1 \mid y = 1)$
versus $P(\alpha_T = 1 \mid y = -1)$

$$P(\alpha_T = 1 \mid y = 1) = \sum_{\underline{x}} P(\alpha_T(\underline{x}) = 1 \mid \underline{x})\, P(\underline{x} \mid y = 1)$$
$$= \sum_{\underline{x} :\, \log \frac{P(\underline{x} \mid y=1)}{P(\underline{x} \mid y=-1)} > T} P(\underline{x} \mid y = 1)$$

$$P(\alpha_T = -1 \mid y = 1) = \sum_{\underline{x}} P(\alpha_T(\underline{x}) = 1 \mid \underline{x})\, P(\underline{x} \mid y = -1)$$
$$= \sum_{\underline{x} :\, \log \frac{P(\underline{x} \mid y=1)}{P(\underline{x} \mid y=-1)} > T} P(\underline{x} \mid y = -1).$$



Rule $\alpha(x) = 1$ if $\log \frac{P(\underline{x} \mid y=1)}{P(\underline{x} \mid y=-1)} > T$

So if $T = -\infty$, then all data is classified as positive
so $P(\alpha_T = 1 \mid y = -1) = P(\alpha_T = 1 \mid y = 1) = 1$

if $T = \infty$, all data is classified as negative $P(\alpha_T = 1 \mid y = -1) = P(\alpha_T = 1 \mid y = 1) = 0$

Bayes decision is given by a specific point T^* on the curve.

ROC Example

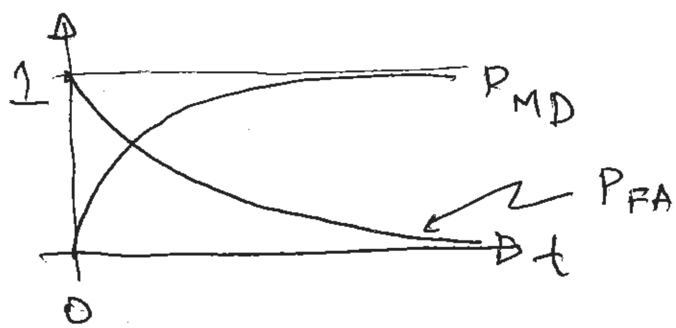
$\rightarrow$ $p(x|y=i) = \frac{1}{\lambda_i} e^{-x/\lambda_i}$, $x \geq 0$ $\quad \lambda_1 \geq \lambda_0$

$\rightarrow$ LRT is equivalent to

$$\hat{y} = \begin{cases} 1 & \text{when } x \geq t \\ 0 & \text{when } x \leq t \end{cases}$$

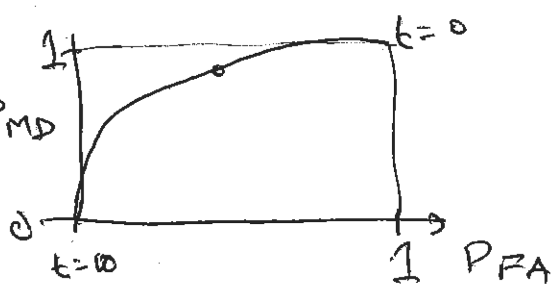
$$P_{FA} = P(\hat{y}=1|y=0) = P(x \geq t | y=0)$$
$$= \int_t^\infty p(x|y=0)\,dx = \frac{1}{\lambda_0}\int_t^\infty e^{-x/\lambda_0}\,dx$$
$$= e^{-t/\lambda_0}$$

$$P_{MD} = P(\hat{y}=0|y=1) = P(x \leq t | y=1)$$
$$= \int_0^t p(x|y=1)\,dx = 1 - e^{-t/\lambda_1}.$$



ROC

$P_D = 1 - P_{MD}$



Optimizing your choice.

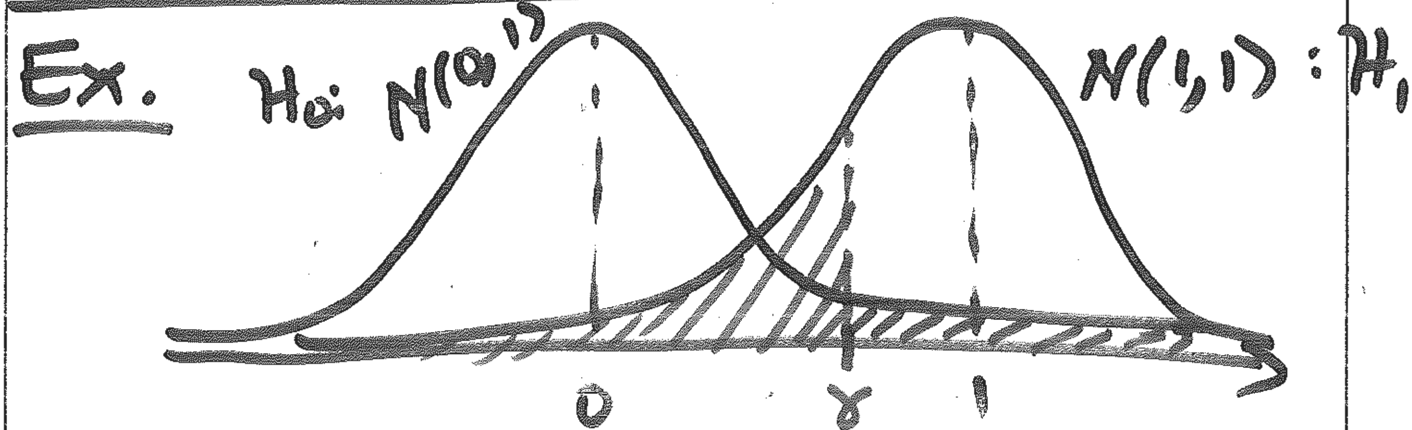
① No prior info

② Prior info. (Bayes Risk)

$Pr(H_0) = p$

$Pr(H_1) = 1-p$

Ex.



$P(x|H_0)$; $p(x|H_1)$

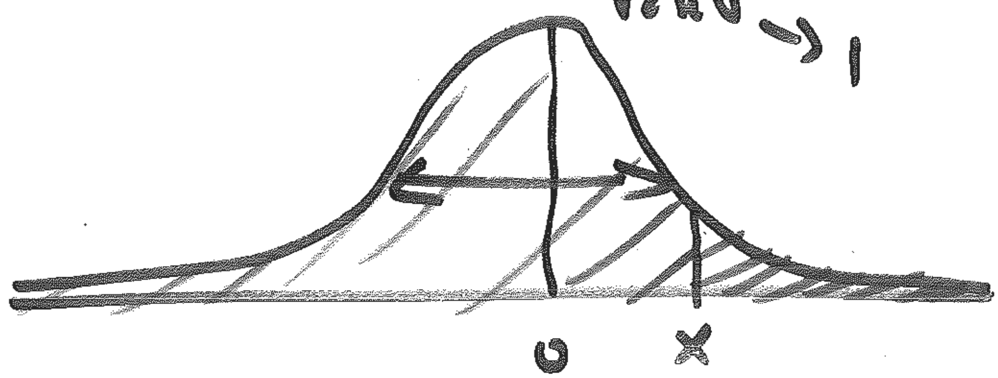
Measure $x \underset{H_1}{\overset{H_0}{\lessgtr}} \gamma$

$Pr(x < \gamma \mid H_1)$

$Pr(x > \gamma \mid H_0)$

A little digression to $\mathcal{N}(0,\sigma^2)$

$\mathcal{N}(0,1)$ $$p(x) = \frac{1}{\sqrt{2\pi}\underbrace{\sigma}_{\to 1}} \exp\left(\frac{x^2}{-2\underbrace{\sigma^2}_{\to 1}}\right)$$



Cumulative dist. fncn.

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \, dt$$

$$Q(x) = 1 - \Phi(x)$$

$$Q(x) \approx \frac{1}{\sqrt{2\pi}\, x} e^{-x^2/2} \qquad x > 4$$

ROC Curver: Receiver Operating Characteristic

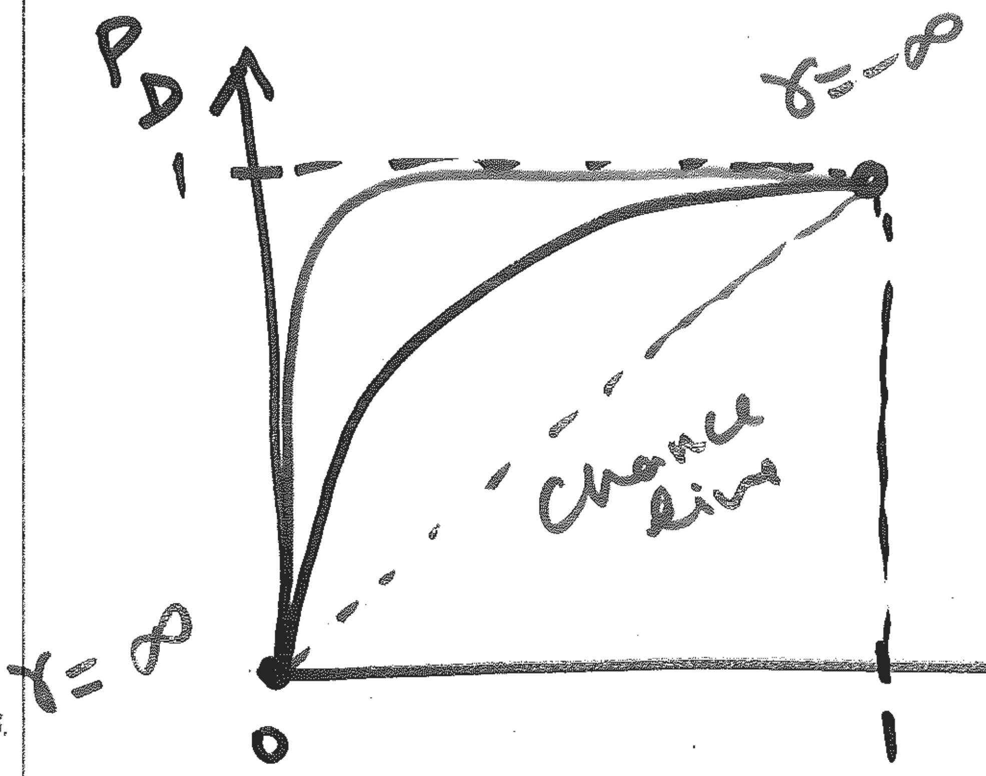
* $Pr(x > \gamma \mid H_0)$ is called the False alarm rate

P_F, P_{FA}

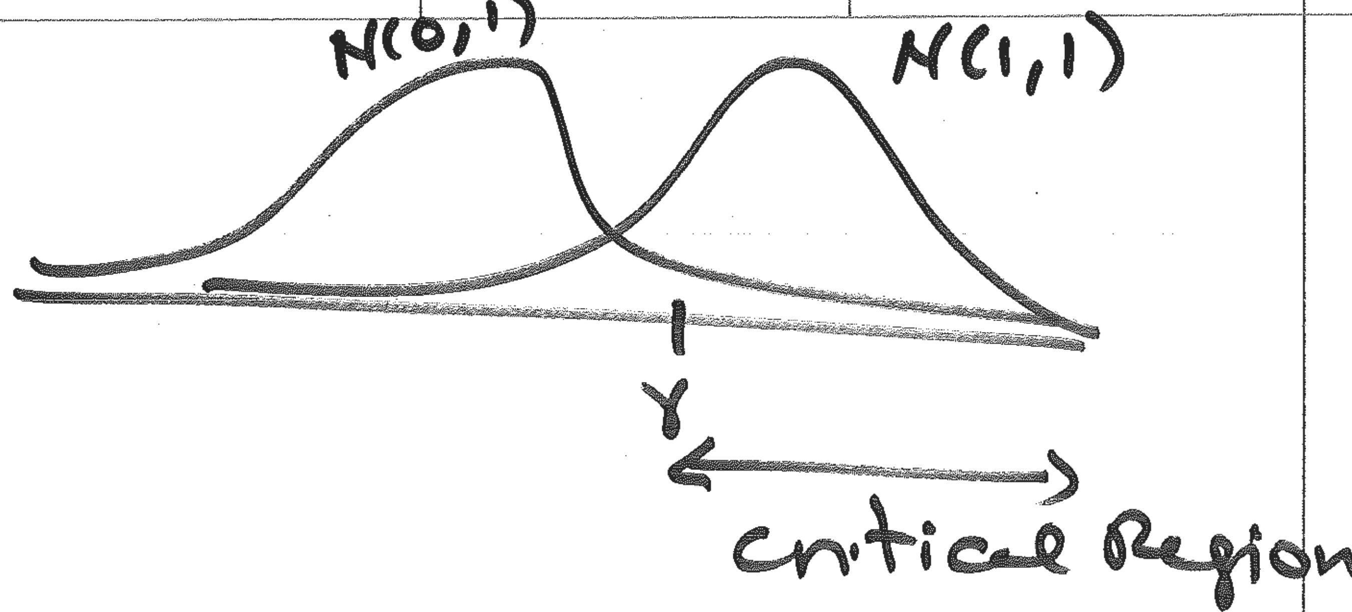
* $Pr(x < \gamma \mid H_1)$ is called missed detection rate (P_M)

$$P_M = 1 - P_D = 1 - \underbrace{Pr(x > \gamma \mid H_1)}$$

The performance is measured or reported using $(P_D, P_F)(\gamma)$



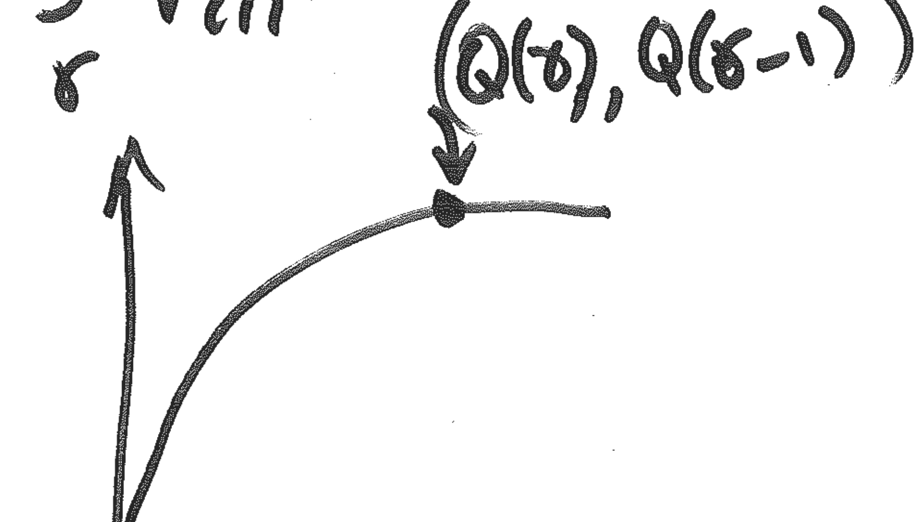
R.O.C.
Receiver
Oper.
Charac.



$$P_F = \Pr(x > \gamma \mid H_0)$$

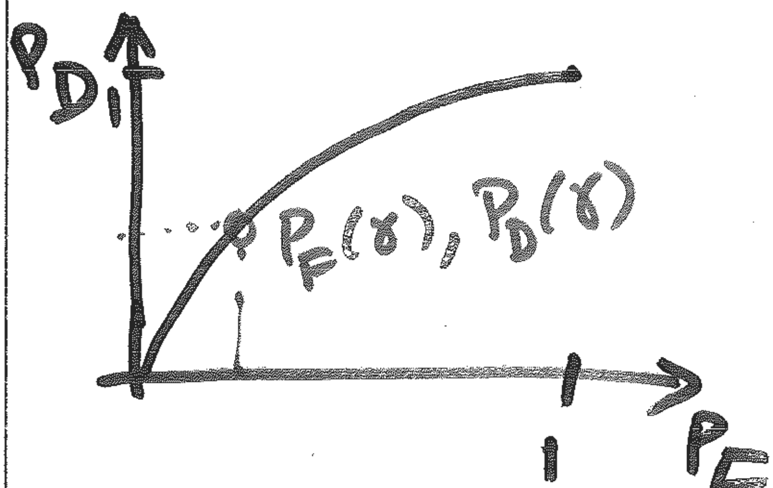
$$= \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt = Q(\gamma)$$

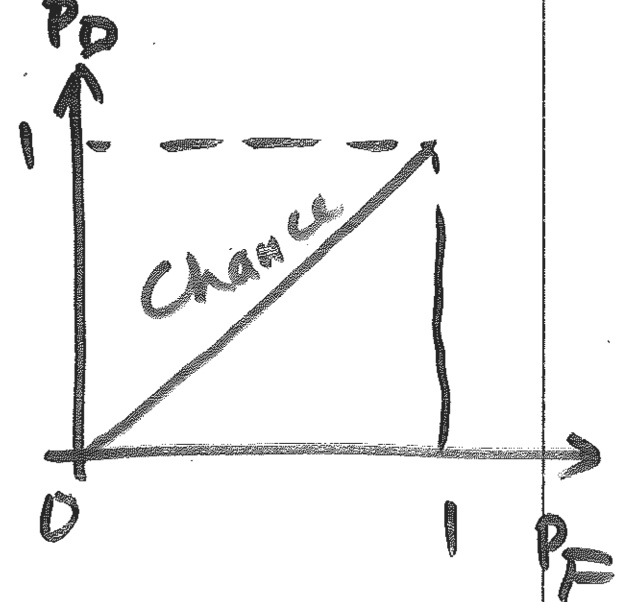
$$P_D = \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(t-1)^2}\, dt = Q(\gamma - 1)$$



TOPS.
35502

ROC curve:





If heads $\rightarrow \mathcal{H}_1$

If tails $\rightarrow H_0$

coin: $Pr(\text{head}) = p$

$$P_F = Pr(\text{heads} \mid H_0) = Pr(\text{head}) = p$$

$$P_D = Pr(\text{heads} \mid H_1) = Pr(\text{heads}) = p$$

Minimum Probability of Error: MAP estimate with loss /cost of 1 for error and 0 for correct decision

Min Prob. of error detection:

Assume knowledge $Pr(H_0)$
$Pr(H_1)$

$$P_e = Pr(\text{pick } H_0 \text{ \underline{and} } H_1 \text{ true}) +$$
$$Pr(\text{pick } H_1 \text{ \underline{and} } H_0 \text{ true})$$

$$= Pr(\text{Pick } H_0 \mid H_1 \text{ true}) \cdot Pr(H_1) +$$
$$Pr(\text{pick } H_1 \mid H_0 \text{ true}) \cdot \underbrace{Pr(H_0)}_{P_0}$$

$$P_e = \underbrace{P_M}_{1-P_D} \cdot (1-P_0) + P_F \cdot P_0$$

$$P_e(\gamma) = P_0 P_F(\gamma) + (1-P_0)\left(1 - P_D(\gamma)\right)$$

MAP: Maximum A Posterior: argmax p(y|x), which is maximizing the probability of being cor
for given evidence so it minimizes the prob kjfjfjfjjfff

Min Prob. Error:

$$\frac{p(x|H_1)}{p(x|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{p_0}{1-p_0}$$

Min Bayes Risk:

$$\underline{\text{Risk}} = \sum_{i,j} C_{ij} \Pr(\text{Choose } H_i | H_j) \cdot \Pr(H_j)$$

R_1 : critical Region $\quad$ R_0 : Complement

$$\text{Risk} = C_{00} \Pr(H_0) \int_{R_0} p(x|H_0)\,dx$$

$$+ C_{01} \Pr(H_1) \int_{R_0} p(x|H_1)\,dx$$

$$+ C_{10} \Pr(H_0) \int_{R_1} p(x|H_0)\,dx$$

$$+ C_{11} \Pr(H_1) \int_{R_1} p(x|H_1)\,dx$$

TOPS 35502

$$\int_{R_0} p(x|H_i)\,dx + \int_{R_1} p(x|H_i)\,dx = 1$$

$$\int_{R_1}\Big[\left[C_{10}\Pr(H_0) - C_{00}\Pr(H_0)\right] p(x|H_0) + \left[C_{11}\Pr(H_1) - C_{01}\Pr(H_1)\right] p(x|H_1)\Big]dx$$

Min Bayes Risk

$$\boxed{\frac{p(x|H_1)}{p(x|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{(C_{10}-C_{00})\Pr(H_0)}{(C_{01}-C_{11})\Pr(H_1)}}$$

<u>Special case</u>

$$C_{00} = C_{11} = 0$$

$$C_{01} = C_{10}$$

$$\boxed{\frac{p(x|H_1)}{p(x|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{\Pr(H_0)}{\Pr(H_1)}}$$ Min Prob. Error

if $\Pr(H_0) = \Pr(H_1) = \frac{1}{2}$

Max Likelihood

$$\boxed{p(x|H_1) \underset{H_0}{\overset{H_1}{\gtrless}} p(x|H_0)}$$

$$\frac{p(x|H_1)}{p(x|H_0)} \gtrless \frac{Pr(H_0)}{Pr(H_1)}$$

$$p(x|H_1)\,Pr(H_1) \gtrless p(x|H_0)\,Pr(H_0)$$

Recall $Pr(H_i|x) = \dfrac{p(x|H_i)Pr(H_i)}{p(x)}$

$$\boxed{Pr(H_1|x) \underset{H_0}{\overset{H_1}{\gtrless}} Pr(H_0|x)}$$

maximum-a-posteriori decision rule

TOPS 35502

(9)

Bayes Decision theory also applies when y is not a binary variable – e.g. y can take M values or y continuous valued

In this course, usually

(i) $y \in \{-1, 1\}$ binary classification.

(ii) $y \in \{1, 2 \ldots M\}$ multi-class classification.

(iii) $y \in (-\infty, \infty)$ regression.

Note: machine learning also addresses cases where $\underline{y} = (y_1, y_2 \ldots, y_N)$ is a vector

$y_i \in \{\pm 1\}$

$y_i \in \{1, 2 \ldots M\}$

$y \in (-\infty, \infty)$

but this is beyond the scope of this course.

(16) <u>Problem</u> (a). Bayes Decision Theory

We usually do not know the distributions $p(y|\underline{x})$ and $p(\underline{x})$

Instead we know data $\mathcal{X}_N = \{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$

<u>E.G.</u> We have bank records of income and savings of N customers, and know if they defaulted or not. $\{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$

income, savings — defaulted or not

<u>Key Assumption</u>

In any Machine Learning problem we assume that the data we observe is generated by an (unknown) probability distribution

The data examples

$$(\underline{x}_1, y_1), (\underline{x}_2, y_2), \ldots, (\underline{x}_N, y_N)$$

are independent, identically distributed (iid) samples from $P(\underline{x}, y)$.

We want to obtain a decision rule $y = \alpha(\underline{x})$ which is good (i.e. small risk) for <u>all</u> possible samples from $p(\underline{x}, y)$ (generalization). A decision rule which has low risk for the data examples (memorization) is not good enough.

(11) This suggests two strategies.

Strategy (1): The Probabilistic Approach.

Use the data $\{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$ to learn probability distributions $P(\underline{x}|y)$ and $p(y)$. Then apply Bayes Decision Theory.

E.G.

$$p(y=1) = \frac{\sum_{i=1}^{N} I(y_i=1)}{N} \qquad p(y=-1) = \frac{\sum_{i=1}^{N} I(y_i=-1)}{N}$$

Indicator function:
$I(y=1) = 1$, if $y=1$
$I(y=1) = 0$, otherwise

E.G. Gaussian assumption $P(\underline{x}|y=1) = N(\underline{\mu}_1, \underline{\Sigma}_1)$, $P(\underline{x}|y=-1) = N(\underline{\mu}_{-1}, \underline{\Sigma}_{-1})$

$N(\cdot, \cdot)$ normal, mean, covariance

$$\underline{\mu}_1 = \frac{\sum_{i=1}^{N} I(y_i=1)\underline{x}_i}{\sum_{i=1}^{N} I(y_i=1)}, \quad \underline{\mu}_{-1} = \frac{\sum_{i=1}^{N} I(y_i=-1)\underline{x}_i}{\sum_{i=1}^{N} I(y_i=-1)}$$

$$\underline{\Sigma}_1 = \frac{1}{\sum_{i=1}^{N} I(y_i=1)} \sum_{i=1}^{N} I(y_i=1)(\underline{x}_i-\underline{\mu}_1)(\underline{x}_i-\underline{\mu}_1)^T$$

$$\underline{\Sigma}_2 = \frac{1}{\sum_{i=1}^{N} I(y_i=-1)} \sum_{i=1}^{N} I(y_i=-1)(\underline{x}_i-\underline{\mu}_{-1})(\underline{x}_i-\underline{\mu}_{-1})^T$$

i.e. estimate the mean and covariances for classes $y=1$ and $y=-1$ using only the data assigned to that class (e.g. assign $\underline{x}_i$ to class $y=1$, if $y_i=1$).

Note: This strategy requires learning parametric and non-parametric probability distributions — we will discuss methods for doing this in later lectures.

(12) Stategy (2): Decision Rule

Learn the decision rule $y = \alpha(\underline{x})$ directly.

Define the empirical risk $R_{emp}(\alpha, \mathcal{X}_N) = \frac{1}{N}\sum_{i=1}^{N} L(\alpha(\underline{x}_i), y_i)$.

This depends on the dataset: $\mathcal{X}_N = \{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$

For example (but not always – see later lectures).

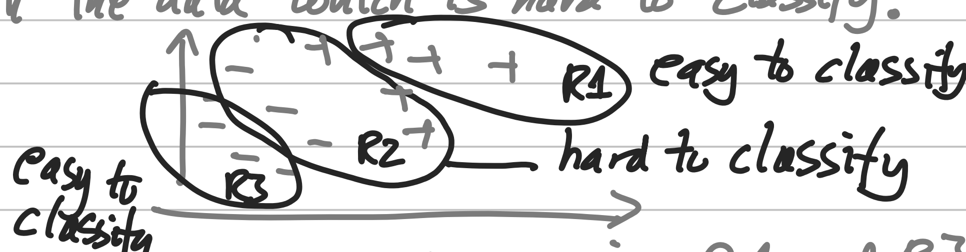
select $\hat{\alpha}(\,) = \arg\min_{\alpha} R_{emp}(\alpha : \mathcal{X}_N)$

Motivation: (i) why bother to learn the probability distributions if your final goal is to obtain the decision rule?

(ii) you may make mistakes when you learn the probability distributions – because you will have to make assumptions about them (see next lectures) which may be incorrect. (see example on next page)

(iii) you should concentrate your effort by dealing with the data which is hard to classify.

Example:

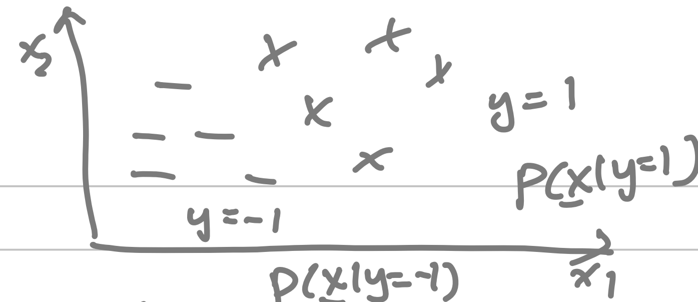


It is easy to classify the data in regions R1 and R3 (because these regions contain only +ve and -ve examples respectively)

So we should concentrate our effort in R2 (+ve and -ve examples)

But the probability strategy would pay equal attention to R1, R2, R3

(13) Danger of Using Probabilities



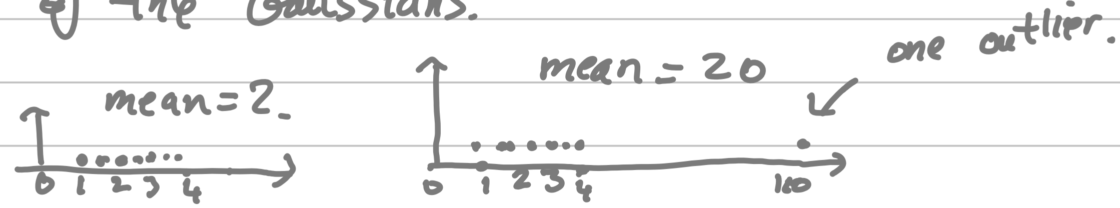
Suppose we assume the distribution $P(\underline{x}|y)$ are Gaussians

But Gaussians are <u>non-robust</u>.

Outliers in the data — unlikely values of $\underline{x}$. — can make big changes to the estimates of the means and covariances of the Gaussians.

<u>Example:</u>

Data.

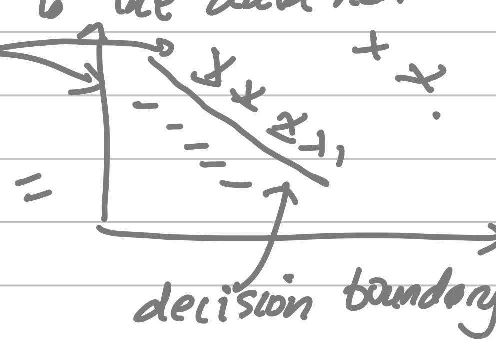


<u>Note</u>: we have to assume a model for $p(\underline{x}|y)$ when we learn from data - if model is wrong, then mistakes happen

So if we try to learn $P(\underline{x}|y)$ our estimates of the mean and covariance, hence of the decision boundary, can be corrupted by outliers far away from the boundary.

But if instead, we just search for a linear plane that seperates the data from $y=1$ and $y=-1$ then we only need to pay attention to the data near the boundary

This is particularly helpful if we have little data, as is usually the case.



(14) Fundamental Problem of Machine Learning

We want to find $\hat{\alpha}(.)$ to minimize Bayes Risk $R(\alpha)$
— this is generalization.

But we only know the empirical risk $R_{emp}(\alpha : \mathcal{X}_N)$
of the dataset $\mathcal{X}_N = \{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$.

Finding $\hat{\alpha}(.)$ to minimize $R_{emp}(\alpha : \mathcal{X}_N)$ may only memorize the dataset, which is not what we want.

Fundamental Assumption: the dataset $\mathcal{X}_N$ consists of i.i.d. samples from $P(\underline{x}, y)$.

Insight. As the dataset gets bigger $N \to \infty$, the empirical risk convergences (in probability) to the Bayes risk. i.e. $R_{emp}(\alpha; \mathcal{X}_N) \to R(\alpha)$

Let A be the set of all decision rules, (e.g. ML, MAP, seperating planes, nearest neighbor, decision trees).

Now suppose N is large enough so that
$|R_{emp}(\alpha : \mathcal{X}_N) - R(\alpha)|$ is small, for all $\alpha \in A$

then we can select a rule $\hat{\alpha} = \underset{\alpha \in A}{\arg\min}\, R(\alpha, N)$

and be confident that

$$R(\hat{\alpha}) \text{ is close to } \min_{\alpha \in A} R(\alpha)$$

i.e. that the rule $\hat{\alpha}$ works well on all data from $P(\underline{x}, y)$, that it generalizes.

How big should N be? It depends on the set A of decision rules. This is Advanced Material.

(15) **Memorization vs. Generalization** Advanced Material

$$R_{emp}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} L(\alpha(x_i), y_i)$$

suppose $L(\alpha(x_i), y_i) \in \{1, 0\}$

By law of large numbers $R_{emp}(\alpha) \underset{N \to \infty}{\longrightarrow} R(\alpha) = \sum_{x,y} p(x,y) L(\alpha(x), y)$

but how fast?

Fix α: By standard theorems (Chernoff, Sanov, Cremers.)

$$\Pr\{|R_{emp}(\alpha) - R(\alpha)| > \epsilon\} < e^{-N\epsilon}$$

require $e^{-N\epsilon} < \delta \iff N > -\frac{1}{\epsilon}\log\delta$ {any ϵ} {$-\log\delta > 0$ if $0 \le \delta \le 1$}

So, if $N > -\frac{1}{\epsilon}\log\delta$, then with prob $> 1-\delta$

$|R_{emp}(\alpha) - R(\alpha)| < \epsilon$ → Almost certain we can estimate the risk of α from $N > -\frac{1}{\epsilon}\log\delta$ examples

Probably Approximately Correct (PAC)

But, we must consider many different rules $\alpha \in A$

For simplicity, suppose we consider a finite no. of rules $\{\alpha^\nu : \nu = 1 \text{ to } H\}$

We want $|R_{emp}(\alpha^\nu) - R(\alpha^\nu)| < \epsilon$ to be small for all ν, with high probability.

Boole's Inequality: $\Pr(A^1 \text{ or } \dots \text{ or } A^H) \le \sum_{\nu=1}^{H} \Pr(A^\nu)$

Let $\Pr(A^\nu)$ be prob that $|R_{emp}(\alpha^\nu) - R(\alpha^\nu)| > \epsilon$

$\Pr\{\text{At least one rule } A^\nu \text{ has error greater than } \epsilon\}$

$< He^{-N\epsilon}$ Now want $He^{-N\epsilon} < \delta$

$\iff N > \frac{1}{\epsilon}\{\log H - \log\delta\}$

So if $N > \frac{1}{\epsilon}\{\log H - \log\delta\}$, then with prob $> 1-\delta$

$|R_{emp}(\alpha^\nu) - R(\alpha^\nu)| < \epsilon$ for all $\nu = 1 \text{ to } H$.

Hence number of examples needed grows rapidly with H (size of hypothesis space), accuracy required ϵ, certainty δ.

(16) <u>Memorization:</u>

Decision Rule: $\hat{\alpha} = \underset{\alpha}{\text{ARGMIN}}\ R_{emp}(\alpha)$

$R_{emp}(\hat{\alpha})$ small, but $R(\alpha)$ big.

i.e. bad for predicting new data.

<u>Generalization:</u>

<u>Want</u> a decision rule $\bar{\alpha}$ so that $R_{emp}(\bar{\alpha})$ is small, but $R(\bar{\alpha})$ is small.

<u>In</u> practice — cross-validation.

training set $\{(x_i, y_i) : i = 1 \text{ to } N\}$ to learn the rule $\bar{\alpha}$

test set $\{(x_j, y_j) : j = 1 \text{ to } M\}$ to test the rule $\bar{\alpha}$.

Choose $\bar{\alpha}$ so that $R_{emp}(\bar{\alpha})$ is small on both the training set and test set.

<u>How</u>, restrict the possibilities of $\bar{\alpha}$.