Chapter 2: Fundamentals of prediction aka Decision Theory
Patterns, Predictions and Actions by Hardt and Recht

Chapter 1: Pattern recognition; by Bishop

1. Probability Basics

We will introduce the basic concepts of probability theory by considering a simple example. Imagine we have two boxes, one red and one blue, and in the red box we have 2 apples and 6 oranges, and in the blue box we have 3 apples and 1 orange. This is illustrated in Figure 1.9. Now suppose we randomly pick one of the boxes and from that box we randomly select an item of fruit, and having observed which sort of fruit it is we replace it in the box from which it came. We could imagine repeating this process many times. Let us suppose that in so doing we pick the red box 40% of the time and we pick the blue box 60% of the time, and that when we remove an item of fruit from a box we are equally likely to select any of the pieces of fruit in the box.

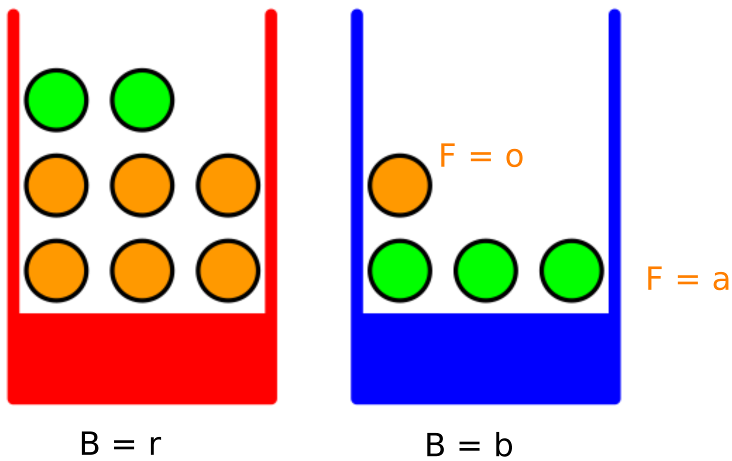


Figure 1.9 We use a simple example of two coloured boxes each containing fruit (apples shown in green and oranges shown in orange) to introduce the basic ideas of probability.

$p(B = r) = 4/10$ and $p(B = b) = 6/10$

$p(F = a \mid B = r) = 2/8$ and $p(F = o \mid B = r) = 6/8$
$p(F = a \mid B = b) = 3/4$ and $p(F = o \mid B = b) = 1/4$

Product rule

Joint prob

$$p(F=a, B=r) = p(F=a \mid B=r)\, p(B=r)$$

sum rule

Marginal prob

$$p(F=a) = \sum_{x \in \{r,b\}} p(F=a, B=x)$$

$$= p(F=a, B=r) + p(F=a, B=b)$$

$$= p(F=a \mid B=r)\,p(B=r) + p(F=a \mid B=b)\,p(B=b)$$

$P(F=a|B=r)$

$\Big\downarrow$ Bayes rule

$P(B=r|F=a)$

$$P(B=r|F=a) = \frac{P(B=r, F=a)}{P(F=a)}$$

$$\boxed{P(B=r|F=a) = \frac{P(F=a|B=r)\,P(B=r)}{P(F=a)}}$$

Bayes rule

The Rules of Probability

sum rule $$p(X) = \sum_Y p(X,Y) \quad (1.10)$$

product rule $$p(X,Y) = p(Y|X)p(X). \quad (1.11)$$

$$P(X) = \sum_Y P(X,Y)$$

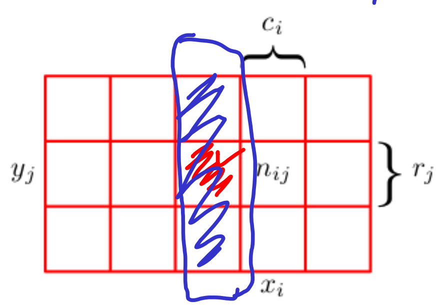


Figure 1.10 We can derive the sum and product rules of probability by considering two random variables, X, which takes the values $\{x_i\}$ where $i = 1, \ldots, M$, and Y, which takes the values $\{y_j\}$ where $j = 1, \ldots, L$. In this illustration we have $M = 5$ and $L = 3$. If we consider a total number N of instances of these variables, then we denote the number of instances where $X = x_i$ and $Y = y_j$ by n_{ij}, which is the number of points in the corresponding cell of the array. The number of points in column i, corresponding to $X = x_i$, is denoted by c_i, and the number of points in row j, corresponding to $Y = y_j$, is denoted by r_j.

$Y = F$

$B = X \rightarrow$ random variable

What is Bayes rule?

Homework 5

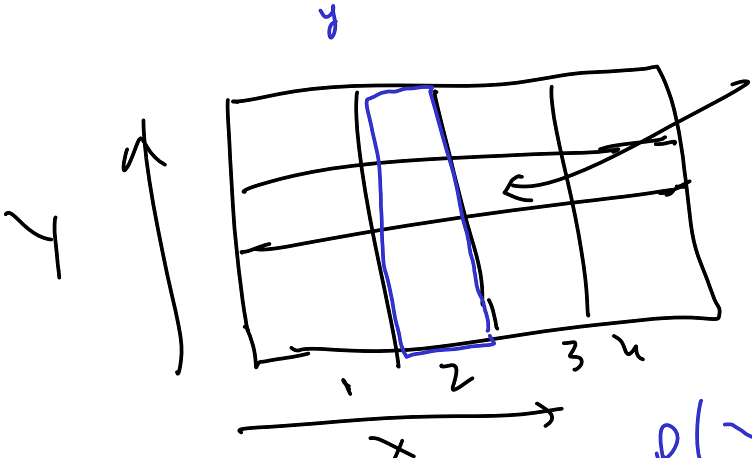
Problem 1

Find the probability of picking the blue box if the chosen fruit was apple?

$$\sum_y P(Y|X=2) = 1$$

Product rule

$P(X,Y) =$



$$\sum_{x,y} P(X,Y) = 1$$

$$P(X=2, Y) = P(Y|X=2)\,P(X=2)$$

Probabilities of continuous variables

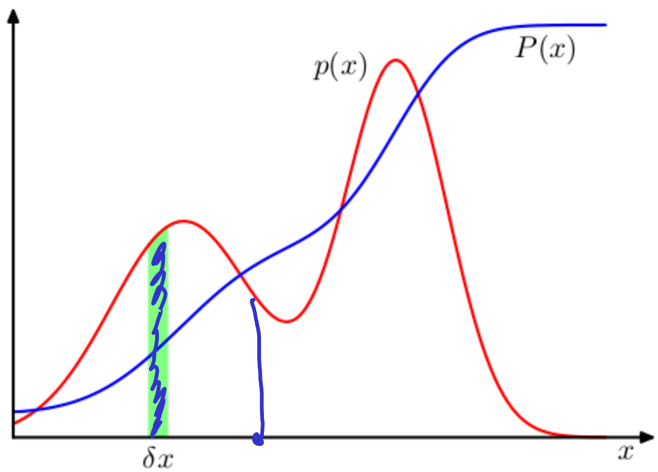


Figure 1.12 The concept of probability for discrete variables can be extended to that of a probability density $p(x)$ over a continuous variable x and is such that the probability of x lying in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$. The probability density can be expressed as the derivative of a cumulative distribution function $P(x)$.

$P(x = 1.7349\ldots) = 0$ $\quad x =$ cont. RV

$P(x \in [1.73, 1.73 + 0.01)) =$ some value

$f(x) =$ Prob density function $= \lim_{\delta x \to 0} \dfrac{P(X \in [x, x + \delta x))}{\delta x}$

$\delta x = 0.01$

$$P(X \in [x, \delta x)) = \int_{x}^{x + \delta x} f(x)\, dx$$

$0 \le f(x)$

$f(x)$ can be > 1

Transformation of random variables

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right| = p_x(g(y))\,|g'(y)|\,.$$

x $y = g(x)$

X is side of square

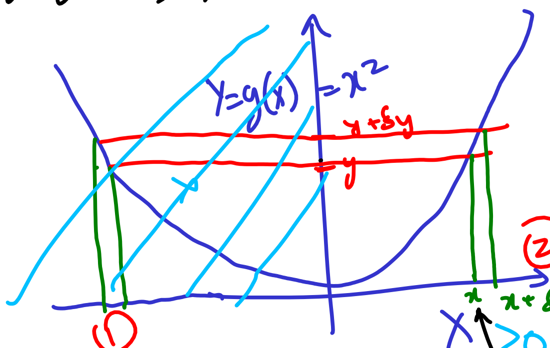
$P_x(X=x) = f_x(x) =$ Prob. density over x

$y = g(x) = x^2$ Transformation

is another RV

Find the PDF of Y random variable

$f_Y(y)\,\delta y = P(Y \in [y, y+\delta y))$



$y = x^2$

$x = \pm\sqrt{y}$

$g^{-1}(y+\delta y)$

$g^{-1}(y)$

$\frac{\delta y}{\delta x} = \frac{d g(x)}{dx} = g'(x) \leftarrow$ derivative of $g(x)$ w.r.t x

$\delta y = g'(x)\,\delta x$ when δx is small

$f_Y(y)\,\delta y = P(Y \in [y, y+\delta y)) = P(X \in [\underbrace{g^{-1}(x)}_{x_g}, \underbrace{g^{-1}(x)}_{x_g} + \underbrace{|g'(x)|\,\delta x}_{\delta x_g}))$

if $g^{-1}(x)$ exist

$= P(X \in [x_g, x_g + \delta x_g))$

LA

1. Capital letters are Matrices
2. bold letters (small) are vectors
3. non bold (small) letters are scalar

Prob.

1. capital letters are RV
2. Small letters are the values RV can take
3. Bold letters denote Random vectors

$$= P\left(X \in [x_g, x_g + |g'(x)|\delta x) \right)$$

$$= g'(x)\, P\left(X \in [x_g, x_g + \delta x) \right)$$

$$f_Y(x) \qquad = g'(x)\, f_X(x_g)\, \delta x$$

$g'(x)$

$x+\delta x$

$g'(x)\,\delta x$

$$P\left(Y \in [y, y+\delta y) \right)$$

$$Y = g(X) = X^2$$

$$= P\left(\underbrace{g^{-1}(Y)}_{X} \in [g^{-1}(y), g^{-1}(y+\delta y)) \right)$$

$\underline{f_X(x)}$ given

$f_Y(y) = ?$

$$= P\left(X \in [g^{-1}(y), \underbrace{g^{-1}(y+\delta y)}) \right)$$

$$\boxed{g^{-1}(y+\delta y) \overset{\text{Taylor}}{\underset{\substack{\text{series}\\ \text{expansion}}}{\approx}} g^{-1}(y) + \underbrace{\frac{d}{dy} g^{-1}(y)}\, \delta y}$$

$+ \ldots$ higher order terms

$$\frac{d}{dy} g^{-1}(y) = \lim_{\delta y \to 0} \frac{g^{-1}(y+\delta y) - g^{-1}(y)}{\delta y}$$

$$\lim_{\delta y \to 0} \left(\frac{d}{dy} g^{-1}(y) \right) \delta y = g^{-1}(y+\delta y) - g^{-1}(y)$$

$$\lim_{\delta y \to 0} \underbrace{g^{-1}(y+\delta y) = g^{-1}(y) + \left[\frac{d}{dy} g^{-1}(y)\right] \delta y}$$

first order Taylor series expansion

$$f_Y(y)\,\delta y$$

$$= P(Y \in [y, y+\delta y))$$

$$= P(X \in [g^{-1}(y), \underbrace{g^{-1}(y+\delta y)}))$$

$$= P(X \in [g^{-1}(y), g^{-1}(y) + \underbrace{\frac{d}{dy} g^{-1}(y)\,\delta y}_{\delta x_g}))$$

$$= P(X \in [g^{-1}(y), g^{-1}(y) + \underline{\delta x_g}))$$

$$= f_X(g^{-1}(y))\,\delta x_g$$

$$\boxed{f_Y(y)\,\cancel{\delta y} = f_X(g^{-1}(y))\left(\frac{d}{dy} g^{-1}(y)\right)\cancel{\delta y}}$$

$$\boxed{f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy} g^{-1}(y)\right|}$$

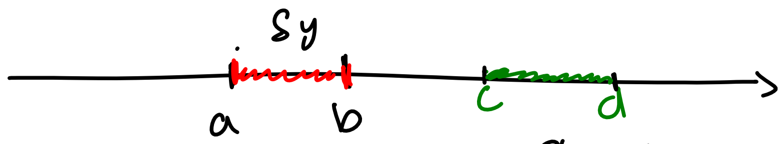
Transformation of cont. RV

Example

$$y = g(x) = x^2 \Rightarrow g^{-1}(y) = +\sqrt{y}$$

$$\Rightarrow \frac{d}{dy} g^{-1}(y) = \frac{1}{2\sqrt{y}}$$

$$f_Y(y) = f_X(\sqrt{y}) \left| \frac{1}{2\sqrt{y}} \right|$$



$x \longrightarrow$ is real

closed interval

$x \in [a, b]$ includes a, b

Open interval

$x \in (c, d)$ excludes c, d

Half closed, half open

$x \in [a, b)$ includes a but excludes b

$x \in (a, b]$ includes b but excludes a

Expectations and covariances

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

$$\mathbb{E}[f] = \int p(x) f(x)\,\mathrm{d}x.$$

A RV X and a function of RV X, f(X)

Discrete RV $\mathbb{E}_X[f] = \sum_{x \in \Omega(X)} P(X=x)\, f(x)$

Cont. RV $\mathbb{E}_X[f] = \int_{x \in \Omega(X)} \underbrace{P(x)}_{\text{Prob density function}} f(x)\, dx$

Conditional expectation

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

6-side dice

$X = 1 \quad 2 \quad 3 \quad 4 \; 5 \; 6$

$P(X) = 1/6$

unbiased dic

$$\mathbb{E}_X[M(X)] = \frac{5}{6} + \frac{15}{6} + 0 + 0 + 0 + 0 = \frac{20}{6} = \frac{10}{3}$$

$$M(X) = \begin{cases} \$5 & \text{if you get } X=2 \\ \$15 & \text{if '' get } X=4 \\ 0 & \text{otherwise} \end{cases}$$

~~Expected value is area under the curve~~

Conditional expectation

$$\mathbb{E}_X[f(X) \mid Y=y] = \sum_x P(X=x \mid Y=y)\, f(x)$$

$$= \int_x p(x \mid Y=y)\, f(x)\, dx$$

Law of iterated expectation

$$\mathbf{E}(X) = \mathbf{E}(\mathbf{E}(X \mid Y)),$$

Sum rule

$$P(X) = \sum_y \underbrace{P(X, Y=y)}$$

$$= \sum_y P(X \mid Y=y) P(Y=y)$$

$$\mathbb{E}_X\left[f(X)\right] = \mathbb{E}_Y\left[\mathbb{E}_X\left[f(X) \mid Y=y\right]\right]$$

$$\underbrace{\mathbb{E}_X\left[f(X) \mid Y=y\right]}_{g(Y)} = \underbrace{\sum_x f(x) P(X=x \mid Y=y)}_{g(Y)}$$

Def of cond expectation

$$\boxed{\underbrace{P(X=x \mid Y=y)}{} = \frac{P(X=x, Y=y)}{P(Y=y)}}$$

$$\mathbb{E}_Y\left[g(Y)\right] = \sum_y g(Y) P(Y=y)$$

$$= \sum_y \left(\sum_x f(x) P(X=x \mid Y=y)\right) P(Y=y)$$

$$= \sum_y \left(\sum_x f(x) P(X=x, Y=y)\right)$$

$$= \sum_x f(x) \sum_y \underbrace{P(X=x, Y=y)}$$

$$= \sum_x f(x) \; P(X=x)$$

$$= \mathbb{E}_X\left[f(X)\right] = LHS$$

$\mu = \mathbb{E}_x[f(x)]$

Variance

$$\mathrm{var}[f] = \mathbb{E}\left[\left(f(x) - \underbrace{\mathbb{E}[f(x)]}\right)^2\right]$$

$$\mathrm{var}_x[f] = \mathbb{E}_x[(f(x) - \mu)^2]$$

Covariance

$$\begin{aligned} \mathrm{cov}[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

$$\mathrm{cov}[x, y] = \mathbb{E}_{x,y}[(x - \mu_x)(y - \mu_y)]$$

$\mu_x = \mathbb{E}_x[x]$

$\mu_y = \mathbb{E}_y[y]$

Covariance in case of random vectors

$$\begin{aligned} \mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]. \end{aligned}$$

Random vector $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

$$\mathrm{cov}(\underline{x}, \underline{y}) = \mathbb{E}_{\underline{x},\underline{y}}\left[\underbrace{(\underline{x} - \underline{\mu}_x)(\underline{y} - \underline{\mu}_y)^T}_{\text{outer product}}\right]$$

$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

Inner product / Dot product

$$\underbrace{\underline{x}^T}_{1\times n} \underbrace{\underline{y}}_{n\times 1} = \underline{x} \cdot \underline{y} = \underline{y}^T \cdot \underline{x} \in \mathbb{R}$$

$$\underline{\mu}_x = \mathbb{E}_{\underline{x}}[\underline{x}] = \begin{bmatrix} \mathbb{E}_{x_1}[x_1] \\ \mathbb{E}_{x_2}[x_2] \\ \vdots \\ \mathbb{E}_{x_n}[x_n] \end{bmatrix}$$

Outer product

$$\underbrace{\underline{x}}_{n\times 1} \underbrace{\underline{y}^T}_{1\times m} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} [y_1 \; \cdots \; y_m]$$

$$= \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_m \\ x_n y_1 & \cdots & \cdots & x_n y_m \end{bmatrix} \in \mathbb{R}^{n \times m}$$

$$\mathrm{cov}(\underline{x}, \underline{y}) = \mathbb{E}_{\underline{x}, \underline{y}} \Big[\underbrace{(\underline{x} - \underline{\mu}_x)(\underline{y} - \underline{\mu}_y)^T}_{\text{outer product}} \Big]$$

$$= \begin{bmatrix} \mathrm{cov}(x_1, y_1) & \cdots & \mathrm{cov}(x_1, y_m) \\ \mathrm{cov}(x_n, y_1) & \cdots & \mathrm{cov}(x_n, y_m) \end{bmatrix}$$

$$\mathrm{cov}(\underline{x}) = \mathrm{var}(\underline{x}) = \mathbb{E}_x \Big[\overbrace{(\underline{x} - \underline{\mu}_x)}^{p} \overbrace{(\underline{x} - \underline{\mu}_x)^T}^{p^T} \Big]$$

$$= \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) & \mathrm{cov}(x_1, x_n) \\ \vdots & \mathrm{var}(x_2) & \\ & & \ddots \\ \mathrm{cov}(x_n, x_1) & & \mathrm{var}(x_n) \end{bmatrix}$$

$\underline{\mathrm{var}(\underline{x})}$ is always positive semi definite

$$\underline{z}^T \mathrm{var}(\underline{x}) \underline{z} \geq 0 \quad \text{for all } \underline{z}$$

outer product

$$\underline{z}^T \boxed{\underline{p}\, \underline{p}^T} \underline{z} = (p^T z)^2 \geq 0$$

Bayes theorem : Posterior, likelihood and prior

weights

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

Dataset

$$\text{posterior} = \frac{\textbf{prior} \times \textbf{likelihood}}{\textbf{evidence}}$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Model = $\hat{y} = f(\underline{x}\,;\,\underline{w})$

prediction — Hypothesis ↳ hidden

input feature

Parameters or weight } Unknowns

$P(D|\underline{w})$ = likelihood of observation

observed

$P(\underline{W})$ = A prior on the thing to be estimated

$P(D)$ = Evidence

$P(\underline{w}|D)$ = Posterior

$$P(D, \underline{W}) = P(\underline{W}, D)$$

$$P(D|\underline{w})\,P(\underline{w}) = P(\underline{w}|D)\,P(D)$$

$$\frac{P(D|w)\,P(\underline{w})}{P(D)} = \underbrace{P(\underline{w}|D)}_{\text{Posterior}}$$

$$\frac{\text{likelihood} \times \text{Prior}}{\text{evidence}} = \text{Posterior}$$

Metrics of binary classification

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	Recall = TPR True positive (TP) TPR = TP/P	False negative (FN) Type II error
	Negative (N)	False positive (FP) Type 1 error	True negative (TN)

Confusion Matrix

Precision = TP/PP

Accuracy = (TP+TN)/(P+N)

(1) Precision : TP / PP
(2) Recall : TP / P
(3) Accuracy : (TP+TN)/(P+N)
(4) F1-score ; Harmonic mean of Precision and Recall

$$\begin{array}{ccc} & \multicolumn{2}{c}{\text{Detected by test}} \\ & \text{cancer} & \text{normal} \end{array}$$

$$\text{True} \quad \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix}$$

loss(1, 1) loss(0, 1)
loss(1, 0) loss(0, 0)

$$\hat{Y} = f(X) = \hat{Y}(X)$$

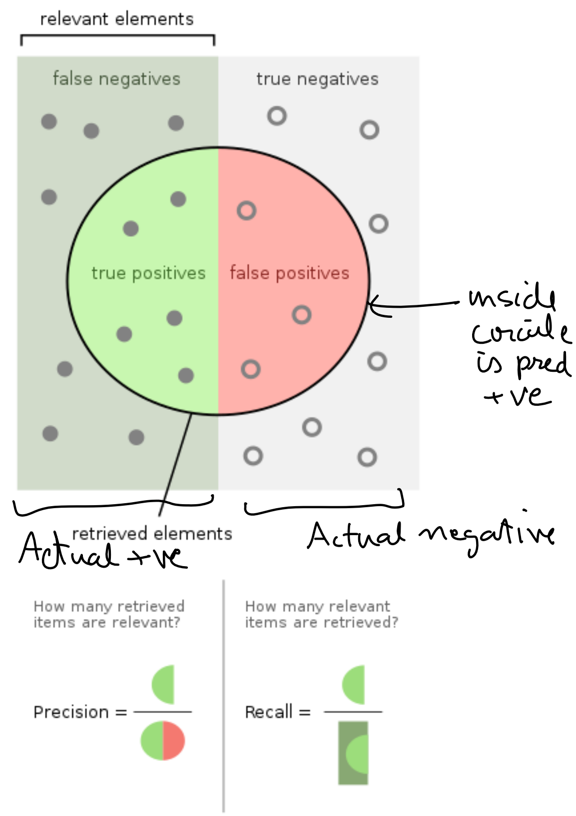
Recall

1. **True Positive Rate:** TPR $= \mathbb{P}[\widehat{Y}(X) = 1 \mid Y = 1]$. Also known as *power*, *sensitivity*, *probability of detection*, or *recall*.
2. **False Negative Rate:** FNR $= 1 -$ TPR. Also known as *type II error* or *probability of missed detection*.
3. **False Positive Rate:** FPR $= \mathbb{P}[\widehat{Y}(X) = 1 \mid Y = 0]$. Also known as *size* or *type I error* or *probability of false alarm*.
4. **True Negative Rate** TNR $= 1 -$ FPR, the probability of declaring $\widehat{Y} = 0$ given $Y = 0$. This is also known as *specificity*.

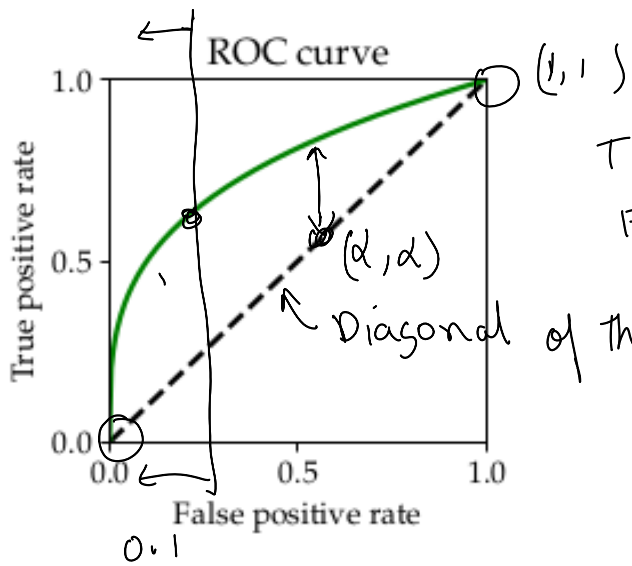
$$TPR = \frac{TP}{P} = \frac{P(\hat{Y}=1, Y=1)}{P(Y=1)} = P(\hat{Y}=1 \mid Y=1)$$

F1-score is the harmonic mean of Precision and Recall

F1-score = 2 * Recall * Precision/(Precision+Recall)



Receiver Operating Characteristics



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

TPR $= P(\hat{Y}(x)=1 \mid Y=1) = \alpha$

FPR $= P(\hat{Y}(x)=1 \mid Y=0) = \alpha$

Diagonal of the ROC curve is a trivial Randomized classifier

Expected risk vs Empirical risk

March 05, 2024

Dataset $D = \{(\underline{x}_1, y_1), \ldots, (x_n, y_n)\}$

images

label

Model $\hat{y}_i = f(\underline{x}_i ; \theta)$

Loss function $\text{loss}(\hat{y}_i, y_i)$

Empirical risk

$$R_D(\theta) = \frac{1}{n} \sum_{i=1}^{n} \text{loss}(\hat{y}_i, y_i)$$

$$\theta^* = \arg\min_{\theta} R_D(\theta)$$

Expected risk

Dataset $\longrightarrow$ Probability distribution

$(\underline{x}_i, y_i)$

Unknown $P_{X,Y}(X, Y)$

Independently
Identically
distributed

i.i.d. assumption

Independently $\equiv$ $(\underline{x}_i, y_i) \perp (\underline{x}_j, y_j)$

$$P\left((\underline{x}_i, y_i) \mid (\underline{x}_j, y_j)\right) = P\left((\underline{x}_i, y_i)\right) \qquad i \neq j$$

$$P\left((\underline{x}_i, y_i), (\underline{x}_j, y_j)\right) = P\left((\underline{x}_i, y_i)\right) P\left((\underline{x}_j, y_j)\right)$$

Identically $\equiv$ $(x_i, y_i) \sim P_{X,Y}(X, Y)$

$\boxed{\text{Expected risk}}$ ← The real target risk we want to minimize

$$R(P_{X,Y}, \theta) = \mathbb{E}_{X,Y}\left[\text{loss}(\hat{Y}, Y)\right]$$

$$= \mathbb{E}_{X,Y}\left[\text{loss}\left(f(X;\theta), Y\right)\right]$$

Training - validation and test split

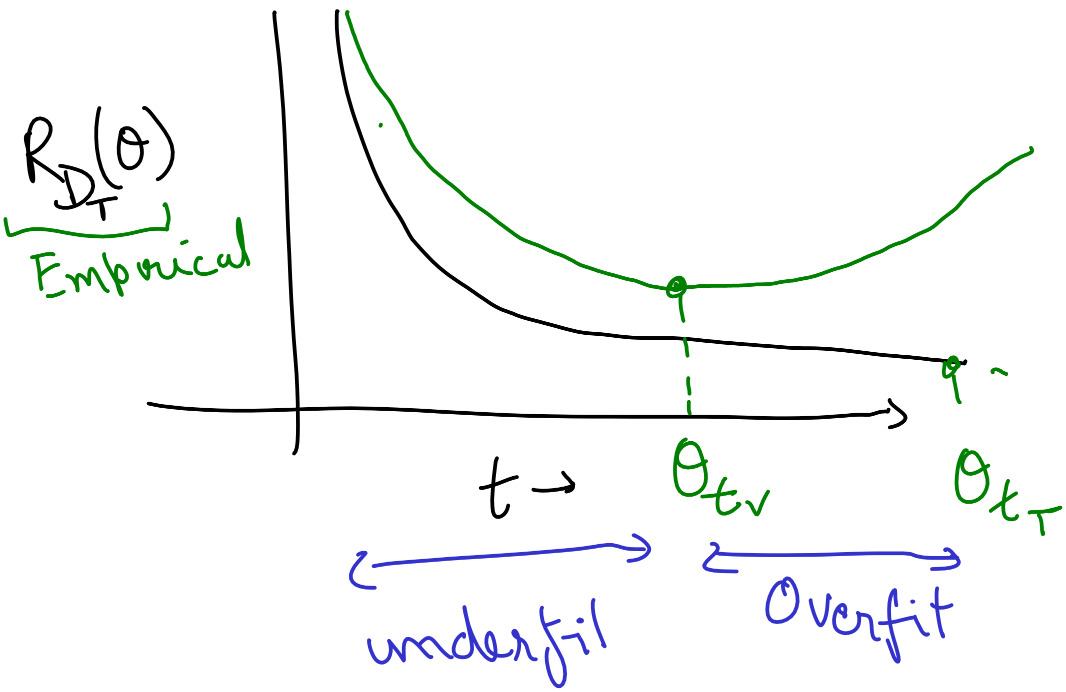
A technique used in ML to bring empirical risk closer to expected risk.

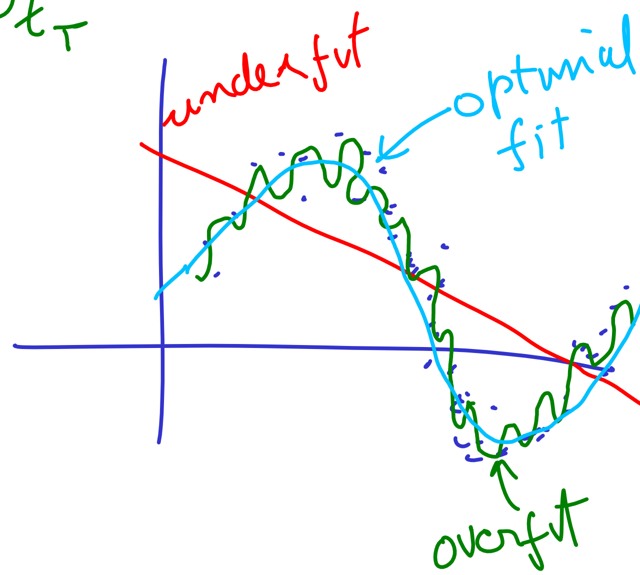
D (10,000)

- Training set (80%) (8,000) $\sim P_{X,Y}$ D_T
- Validation set (10%) (1,000) $\sim P_{X,Y}$ D_V

- Test set (10%) (1,000) $\sim P_{X,Y}$ D_E

Only used for evaluation





Rewrite Bayes rule if the observable is X and hypothesis is the class Y

$$\underbrace{P(Y|X)}_{\text{Posterior}} = \frac{\overbrace{P(X|Y)}^{\text{Likelihood}}\overbrace{P(Y)}^{\text{Prior}}}{\underbrace{P(X)}_{\text{Evidence}}}$$

$$P(\text{Hypothesis}|\text{Obs}) = \frac{\underbrace{P(\text{Obs}|\text{Hypothesis})}_{\text{Likelihood}}\overbrace{P(\text{Hyp})}^{\text{Prior}}}{\underbrace{P(\text{Obs})}_{\text{Evidence}}}$$

Maximum a posteriori predictor

Two class problem $Y \in \{0,1\}$ $\quad \hat{Y}(X) = f(X;\theta)$

$$\hat{Y}(x) = \begin{cases} 1 & \text{if } P(Y=1|X=x) > P(Y=0|X=x) \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{Y}(x) = \arg\max_y P(Y=y|X=x)$$

Introducing loss into optimal decision making

Loss function

		Predicted cancer ($\hat{Y}=1$)	normal ($\hat{Y}=0$)
True (Y)	cancer (1)	0	1000
	normal ($Y=0$)	1	0

$$R = \mathbb{E}_Y\left[\text{loss}(\hat{Y}, Y) \mid X=x\right]$$

$$\hat{Y}(x) = \arg\min_{y} \mathbb{E}_{Y|X}\left[\text{loss}(\hat{Y}, Y=y) \mid X=x\right]$$

Two class problem $Y=\{0,1\}$, $\hat{Y}=\{0,1\}$

Example

$\hookrightarrow \text{loss}(1,1)=0$ $\qquad \text{loss}(0,1)=1000$

$\text{loss}(1,0)=1$ $\qquad \text{loss}(0,0)=0$

Given some observation $X=x$, we have two possible case $\hat{Y}=0$, $\hat{Y}=1$

if $\hat{Y}=0$

RY

$$E_Y\left[\text{loss}(\hat{Y}, Y) \mid X=0\right] = \sum_{y\in\{0,1\}} \text{loss}(\hat{Y}, Y=y)\, P(Y=y \mid X=x)$$

2

$$= \boxed{\text{loss}(0,0)\, P(Y=0 \mid X=x) + \text{loss}(0,1)\, P(Y=1 \mid x)}$$

shortcut

if $\hat{Y}=1$

1

$$\mathbb{E}_Y\left[\text{loss}(\hat{Y}, Y) \mid X=x\right] = \boxed{\text{loss}(1,0)\, P(Y=0 \mid x) + \text{loss}(1,1)\, P(Y=1 \mid x)}$$

$$\hat{y}(x) = \begin{cases} 1 & \text{if } \underbrace{① > ②}_{} \\ 0 & \text{otherwise} \end{cases}$$

① > ② $\quad loss(0,0)\, P(Y=0|x) + loss(0,1)\, P(Y=1|x)$

$\quad > loss(1,0)\, P(Y=0|x) + loss(1,1)\, P(Y=1|x)$

$$\Rightarrow P(Y=1|x)\left[loss(0,1) - loss(1,1)\right] > \left(loss(1,0) - loss(0,0)\right) P(Y=0|x)$$

$$\Rightarrow P(Y=1|x) > \underbrace{\left(\frac{loss(1,0) - \boxed{loss(0,0)}}{loss(0,1) - \boxed{los(1,1)}}\right)}_{} P(Y=0|x)$$

This ratio is the difference between this test and MAP predictor

$$\Rightarrow \frac{P(Y=1|x)}{P(Y=0|x)} > \left(\frac{loss(1,0) - \boxed{loss(0,0)}}{loss(0,1) - los(1,1)}\right) = \varepsilon$$

epsilon

Posterior ratio test

Lemma 1. *We claim that the optimal predictor is given by*

$$\hat{Y}(x) = \mathbb{1}\left\{\mathbb{P}[Y=1 \mid X=x] \geq \frac{loss(1,0) - loss(0,0)}{loss(0,1) - loss(1,1)}\ \mathbb{P}[Y=0 \mid X=x]\right\}$$

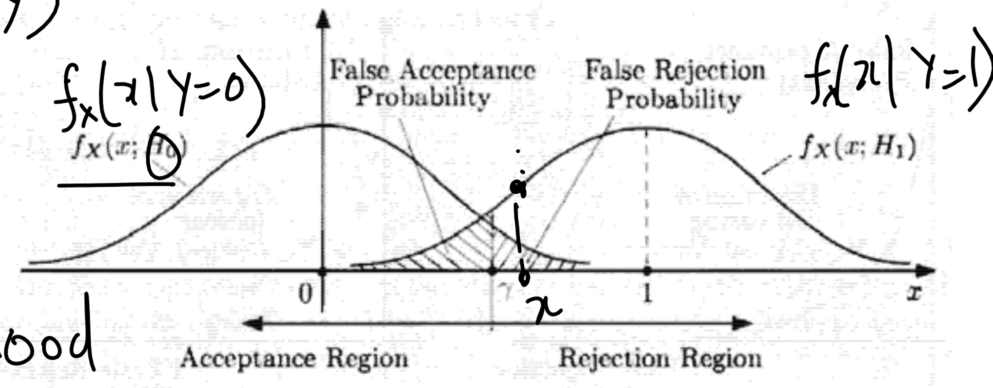
$$\mathbb{1}\{z\} = \begin{cases} 1 & \text{if } z \text{ is true} \\ 0 & \text{if } z \text{ is false} \end{cases}$$

$$\mathbb{E}[loss(0,Y) \mid X=x] = loss(0,0)\,\mathbb{P}[Y=0 \mid X=x] + loss(0,1)\,\mathbb{P}[Y=1 \mid X=x]$$
$$\mathbb{E}[loss(1,Y) \mid X=x] = loss(1,0)\,\mathbb{P}[Y=0 \mid X=x] + loss(1,1)\,\mathbb{P}[Y=1 \mid X=x]\,.$$

Prior X

Likelihood ratio and likelihood ratio test

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$



$$P(X|Y) = \text{likelihood}$$

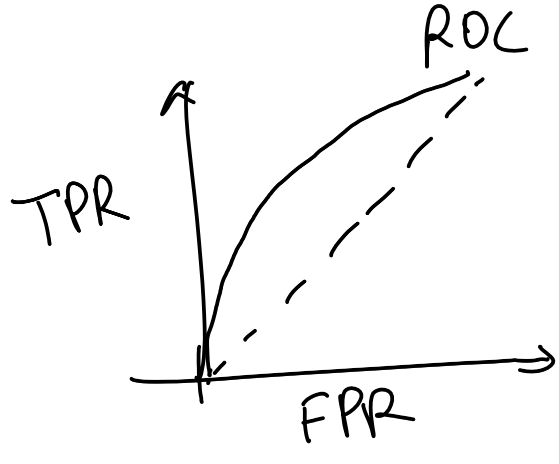
$$\text{Likelihood ratio} = L(x) = \frac{P(X=x \mid Y=1)}{P(X=x \mid Y=0)} \quad \Big| \text{ for DRV } X$$

$$L(x) = \frac{f_X(x \mid Y=1)}{f_X(x \mid Y=0)}$$

Likelihood ratio test

$$\hat{Y}(x) = \begin{cases} 1 & \text{if } L(x) > \eta \\ 0 & \text{otherwise} \end{cases}$$

η ← likelihood test threshold



By varying eta, you can trade off between Precision/Recall. Or between FPR and FNR.

$$\frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} > \varepsilon$$

$$\Rightarrow \frac{P(X=x \mid Y=1)\,P(Y=1)}{\cancel{P(X=x)}} \times \frac{\cancel{P(X=x)}}{P(X=x \mid Y=0)\,P(Y=0)} > \varepsilon$$

$$\Rightarrow L(x)\,\frac{P(Y=1)}{P(Y=0)} > \varepsilon$$

$$\Rightarrow L(x) > \varepsilon\,\frac{P(Y=0)}{P(Y=1)} = \eta$$

Neyman Pearson Lemma

For a given FPR $P(\hat{Y}=1 \mid Y=0) \leq \beta$

$$\hat{Y}(x) = \mathbb{1}\{L(x) > \eta\}$$

gives you the maximum TPR among all predictors $P(\hat{Y}=1 \mid Y=1) = \alpha$

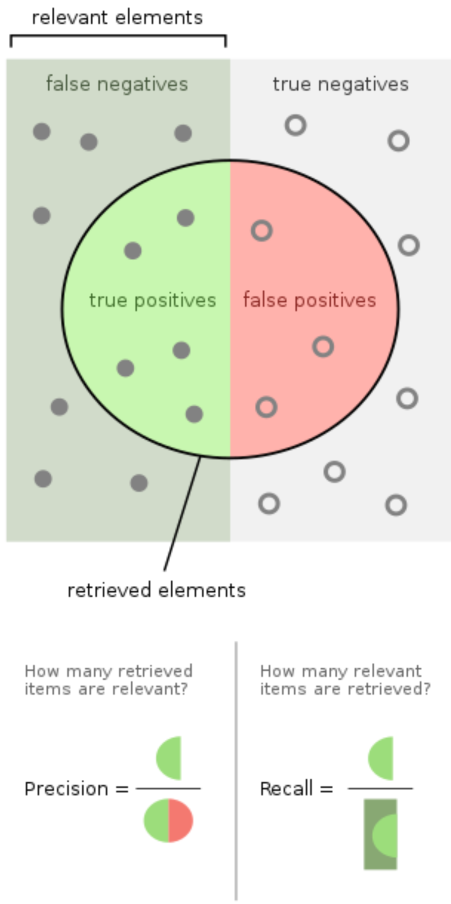
Likelihood Ratio Test (LRT)

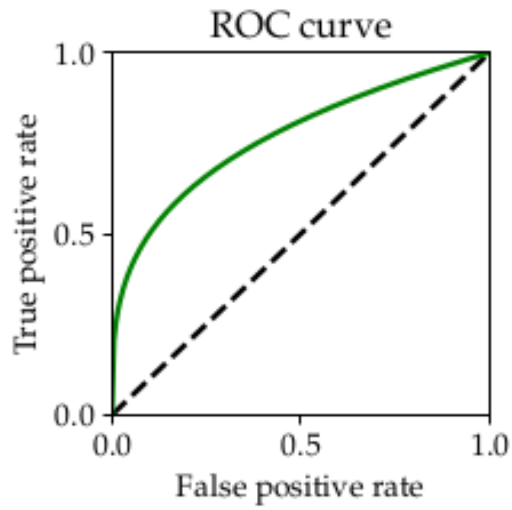
- Start with a target value α for the false rejection probability.
- Choose a value for ξ such that the false rejection probability is equal to α:

$$\mathbf{P}\big(L(X) > \xi; H_0\big) = \alpha.$$

- Once the value x of X is observed, reject H_0 if $L(x) > \xi$.

Neyman-Pearson Lemma





Neyman-Pearson Lemma

Consider a particular choice of ξ in the LRT, which results in error probabilities

$$\mathbf{P}\big(L(X) > \xi; H_0\big) = \alpha, \qquad \mathbf{P}\big(L(X) \leq \xi; H_1\big) = \beta.$$

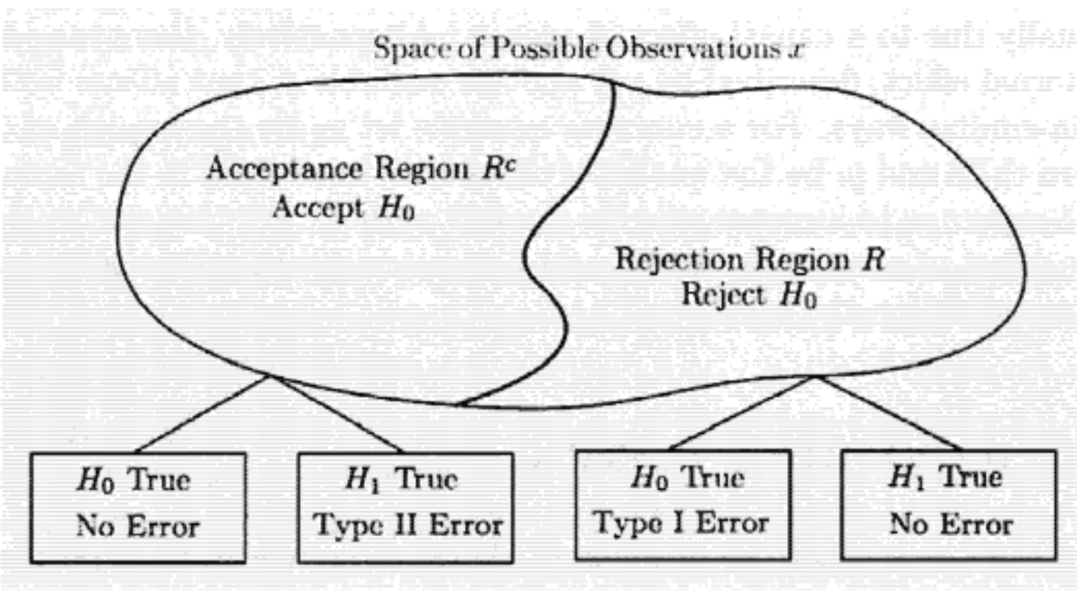
Suppose that some other test, with rejection region R, achieves a smaller or equal false rejection probability:

False positive rate $$\mathbf{P}(X \in R; H_0) \leq \alpha.$$

Then,

False negative rate $$\mathbf{P}(X \notin R; H_1) \geq \beta,$$

with strict inequality $\mathbf{P}(X \notin R; H_1) > \beta$ when $\mathbf{P}(X \in R; H_0) < \alpha$.

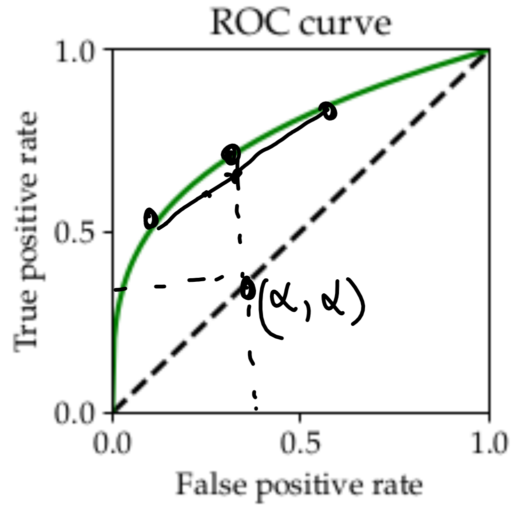


Maximum likelihood predictor

$P(\hat{Y}=1) = \alpha$

$P(\hat{Y}=1 \mid Y=1) = \alpha \Leftrightarrow TPR$

$P(\hat{Y}=1 \mid Y=0) = \alpha \Leftrightarrow FPR$



Proposition 1. The points (0, 0) and (1, 1) are on the ROC curve.

$L(x) > \eta_1$ $\qquad \alpha L(x) + (1-\alpha) L(x) > \alpha \eta_1 + (1-\alpha)\eta_2$

$L(x) > \eta_2$

Proposition 2. The ROC must lie above the main diagonal.

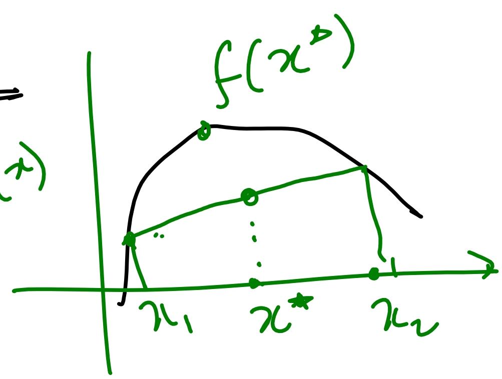
Neyman Pearson Lemma

Likelihood ratio test will always do better

Proposition 3. The ROC curve is concave.

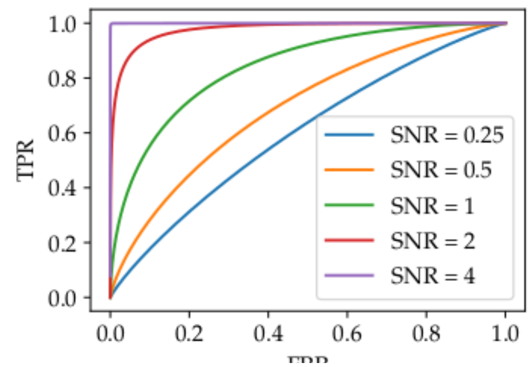
$f(x^*) > \underbrace{\alpha x_1 + (1-\alpha) x_2}_{x^*}$

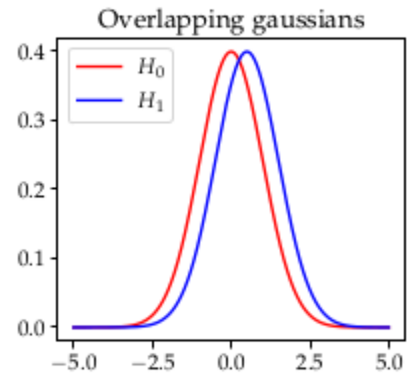
$\alpha \in [0,1]$



$f(\alpha x_1 + (1-\alpha) x_2) > \alpha x_1 + (1-\alpha) x_2 \Leftrightarrow f(x)$ is concave for all x_1, x_2, $\alpha \in [0,1]$

$f(\alpha x_1 + (1-\alpha) x_2) < \alpha x_1 + (1-\alpha) x_2 \Leftrightarrow f(x)$ is convex


TPR
FPR
SNR = 0.25
SNR = 0.5
SNR = 1
SNR = 2
SNR = 4


Overlapping gaussians
H_0
H_1

Example 9.10. We have a six-sided die that we want to test for fairness, and we formulate two hypotheses for the probabilities of the six faces:

$$H_0 \text{ (fair die):} \qquad p_X(x; H_0) = \frac{1}{6}, \qquad x = 1, \ldots, 6,$$

$$H_1 \text{ (loaded die):} \qquad p_X(x; H_1) = \begin{cases} \frac{1}{4}, & \text{if } x = 1, 2, \\ \frac{1}{8}, & \text{if } x = 3, 4, 5, 6. \end{cases}$$

Find a likelihood ratio test for discriminating between a fair die and a loaded die . Also determine True positive rate and false positive rate as a function of likelihood ratio threshold.

Consider, for example, a medical diagnosis problem in which we have taken an X-ray image of a patient, and we wish to determine whether the patient has cancer or not. In this case, the input vector x is the set of pixel intensities in the image, and output variable t will represent the presence of cancer, which we denote by the class C1 , or the absence of cancer, which we denote by the class C2. We might, for instance, choose t to be a binary variable such that t = 0 corresponds to class C1 and t = 1 corresponds to class C2. The general inference problem then involves determining the joint distribution p(x, Ck), or equivalently p(x, t), which gives us the most complete probabilistic description of the situation.

We would like this choice to be optimal in some appropriate sense.
This is the decision step, and it is the subject of decision theory to tell us how to make optimal decisions given the appropriate probabilities.

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}.$$

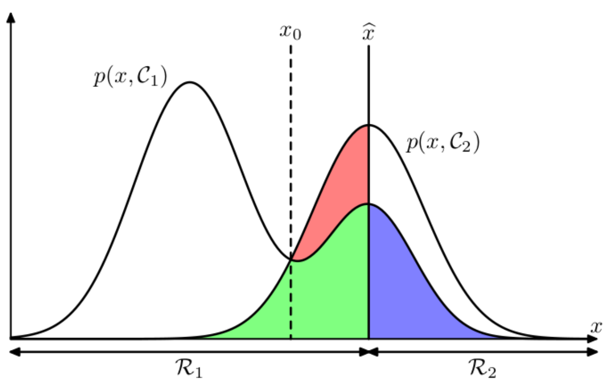


Figure 1.24 Schematic illustration of the joint probabilities $p(x, \mathcal{C}_k)$ for each of two classes plotted against x, together with the decision boundary $x = \widehat{x}$. Values of $x \geqslant \widehat{x}$ are classified as class $\mathcal{C}_2$ and hence belong to decision region $\mathcal{R}_2$, whereas points $x < \widehat{x}$ are classified as $\mathcal{C}_1$ and belong to $\mathcal{R}_1$. Errors arise from the blue, green, and red regions, so that for $x < \widehat{x}$ the errors are due to points from class $\mathcal{C}_2$ being misclassified as $\mathcal{C}_1$ (represented by the sum of the red and green regions), and conversely for points in the region $x \geqslant \widehat{x}$ the errors are due to points from class $\mathcal{C}_1$ being misclassified as $\mathcal{C}_2$ (represented by the blue region). As we vary the location $\widehat{x}$ of the decision boundary, the combined areas of the blue and green regions remains constant, whereas the size of the red region varies. The optimal choice for $\widehat{x}$ is where the curves for $p(x, \mathcal{C}_1)$ and $p(x, \mathcal{C}_2)$ cross, corresponding to $\widehat{x} = x_0$, because in this case the red region disappears. This is equivalent to the minimum misclassification rate decision rule, which assigns each value of x to the class having the higher posterior probability $p(\mathcal{C}_k|x)$.

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}. \end{aligned}$$

$$
\begin{aligned}
p(\text{correct}) &= \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) \\
&= \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}
\end{aligned}
$$

ROC curves