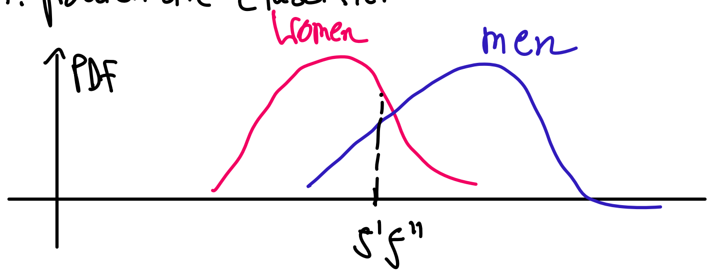
Decision Theory (risk minimization)

adult height → man/woman

1st issue: multiple pts w/ different classes could lie at same pt

Want: probabilistic classifier



Suppose 10% of population has cancer, 90% doesn't

Prob distributions for calorie intake, $P(X|Y)$: X = # calories Y = cancer

Calories (X)	$\le 1{,}200$	1200–1600	> 1600
Cancer ($y=1$)	20%	50%	30%
$\overline{\text{Cancer}}$ ($y=-1$)	1%	10%	89%

Recall: $P(x) = P(x|y)P(y) + P(x|\bar{y})P(\bar{y})$

$$P(1200 \le X \le 1600) = 0.5 \cdot 0.1 + 0.9 \cdot 0.10 = 0.14$$

guy eats $X = 1400$ calories. prob(guy has cancer)

bayes rule (posterior probability) $= \dfrac{\text{prob}(12\text{-}1600 \mid \text{cancer})\overbrace{P(\text{cancer})}^{\text{prior probability}}}{P(1200 \le x \le 1600)} = \dfrac{0.2 \cdot 0.1}{0.14} = \dfrac{0.05}{0.14}$

*assumes equal loss for misclassifications

↳ need to punish false negatives more

$$\rightarrow P(\text{cancer} \mid 1200 \le x \le 1600) = \frac{5}{14} \approx 36\%$$

loss function $L(z, y)$ specifies badness if classifier predicts z but classifies y

e.g. $L(z, y) = \begin{cases} 1 & z = 1,\ y = -1 \\ 5 & z = -1,\ y = 1 \\ 0 & z = y \end{cases}$

false positive is bad

false negative really bad

correct classification good

asymmetrical

36% probability of loss 5, worse than 64% chance of loss 1

get next test/biopsy

Definitions:

Symmetric loss fn: penalties are equal

0-1 loss fun: 0 if True, 1 if false/incorrect

let $r: \mathbb{R}^d \to \pm 1$ (decision rule/classifier)

↳ maps feature vector to prediction

Risk for a decision rule R = expected loss over all possible values of x & y

Risk = expected loss

$$\rightarrow R(r) = E[L(r(X), Y)]$$

functional

posterior

$$= \sum_{x} \Big[\big(\underbrace{L(r(x), 1)}_{\text{outcome}} \cdot \underbrace{P(Y=1|X)}_{\text{prob of outcome}} + L(r(x), -1) P(Y=-1|X=x) \Big] P(X=x)$$

set of all possible values of RV X

prior

Points not in class

$$\text{(Baye's Rule)} = P(Y=1) \sum_{x} L(r(x),1) P(X=x|Y=1) + \overbrace{P(Y=-1) \sum_{x} L(r(x),-1) P(X=x|Y=-1)}$$

Bayes decision rule: function r^* that minimizes functional $R(r)$ (takes in a function)
(Bayes Classifier)

Assuming $L(z, y) = 0$ for $z = y$:

discrete case

$$r^*(x) = \begin{cases} 1 & \text{if } L(-1,1) P(Y=1|X=x) > L(1,-1) P(Y=-1|X=x) \\ -1 & \text{O/W} \end{cases}$$

(which of 2 terms is bigger)

When L is symmetric, $L(-1,1) = L(1,-1) \Longrightarrow$ only compare $P(Y=1|X=x) > P(Y=-1|X=x)$

↳ In other words: Pick class w/ biggest posterior probability

If not symmetric: weigh posteriors by respective loss functions

for cancer example: $r^*(x) = 1$ for $x \le 1600$, $r^*(x) = -1$ for $x > 1600$

he chose these

Bayes/optimal risk: risk of best possible classifier (Bayes classifier)

Cancer example: $R(r^*) = 0.1(5\cdot 0.3) + 0.9(1\times 0.01 + 1\cdot 0.1) = \underbrace{0.249}_{\text{no decision rule gives lower risk}}$

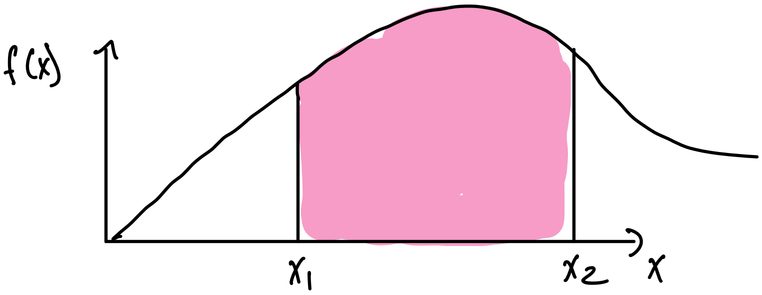
* using prior prob. formula

Deriving/using r^* is called risk minimization

CONTINUOUS DISTRIBUTIONS

X is continuous w/ a PDF

review:



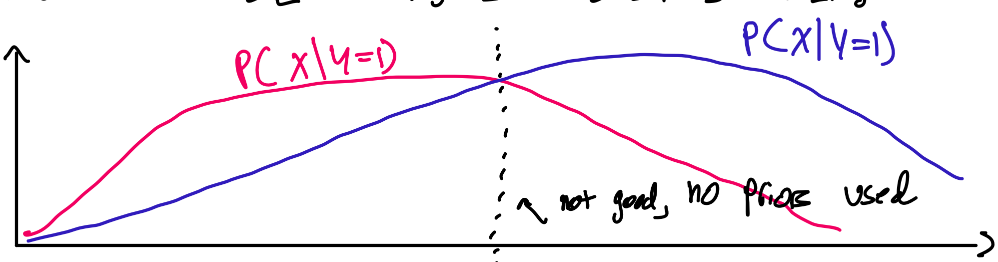
prob that $X \in [x_1, x_2] = \int_{x_1}^{x_2} f(x)\,dx$

$x \in [-\infty, \infty] = 1$ (def of pdf) $= \int_{-\infty}^{\infty} f(x)\,dx = 1$

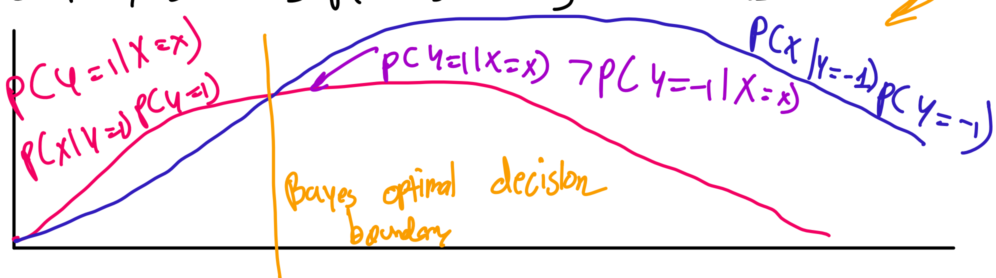
Expected value of $g(X) = \int_{-\infty}^{\infty} g(x) f(x)\,dx$

mean $\mu = E[X] = \int_{-\infty}^{\infty} x f(x)\,dx$

Variance: $\sigma^2 = E[(X-\mu)^2] = E[X^2] - E[X]^2$



Suppose $P(Y=1) = 1/3$, $P(Y=-1) = 2/3$, 0-1 loss



Posterior probabilities, don't need to divide by $P(X)$

Define risk as before, Integral instead of Summation

$$R(r) = E[L(r(x), Y)]$$
$$= P(Y=1)\int L(r(x), 1)\, f(X=x \mid Y=1)\, dx$$
$$+ P(Y=-1)\int L(r(x), -1)\, f(X=x \mid Y=-1)\, dx$$

Bayes risk = area under minimum of 2 functions

Assuming $L(z,y)=0$ for $z=y$:

$$R(r^*) = \int \min_{y=\pm 1} L(-y,y)\, f(X=x \mid Y=y) P(Y=y)\, dx$$

0-1 loss $\Longrightarrow$ Risk $= P(r(x) \text{ is wrong})$

posterior probability

$\hookrightarrow$ Bayes optimal decision boundary $= \{x : \underbrace{P(Y=1 \mid X=x)}_{\text{decision function}} = \underset{\text{isovalue}}{0.5}\}$

3 WAYS TO BUILD Classifiers

1. generative models (e.g. LDA)
   - Assume sample pts come from probability distributions, different for each class
   - guess form of distributions
   - For each class C_i, fit distribution parameters to the class, giving $f(X \mid Y=C_i)$
   - estimate prior probability for each class $P(Y=c)$
   - get posterior probability w/ Bayes thm
   - If 0-1, pick C that maximizes $P(Y=c \mid X=x)$ equivalent to maximizing $f(X=x \mid Y=C)P(Y=c)$

2. Discriminative Models (e.g. logistic regression)
   - model posterior probability w/o class conditional probabilities

3. Find decision boundary (e.g. SVM)

   - model $r(x)$ directly, (no posterior)

Advantages of 1 & 2: - $P(Y \mid X)$ tells probability you're wrong

advantage of 1: easy to diagnose outliers: $f(x)$ very small

disadvantage of 1: hard to estimate distributions, e.g. real distributions