Lecture 2

STAT161/261 Introduction to Pattern Recognition and Machine Learning

Spring 2019

Prof. Allie Fletcher

Course Admin

- People:
  - Prof. Allie Fletcher.
  - TA: Ruiqi Gao ruiqigao@ucla.edu
- Where:
  - MW 3:30-4:45pm, Public Affairs Bldg 2238
- Grading:
  - C261: Midterm 20%, Final 35%, HW and labs 25%, Quizzes&Participation 10%, Project 10%,
  - C161: Midterm 20%, Final 35%, HW and labs 35%, Quizzes&Participation 10%
  - Project is for graduate students only (see below)
  - Homework will include programming assignments
  - Midterm tentatively May 8
  - Midterm and final are closed book. Equation sheet is provided.

Outline

- Decision Theory
  - Classification, Maximum Likelihood and Log likelihood
  - MAP Estimation, Bayes Risk
  - Probability of errors, ROC
- Empirical Risk Minimization
  - Problems with decision theory, empirical risk minimization
  - Probably approximately correct learning
- Curse of Dimensionality
- Parameter Estimation
  - Probabilistic models for supervised and unsupervised learning
  - ML and MAP estimation
  - Examples

Classification

- How to make decision in the presence of uncertainty?
- History: Prominent in WWII: radar for detecting aircraft, codebreaking, decryption
- Observed data x ∈ X, state y ∈ Y
- p(x | y): conditional distribution

  For each class, model of how the data is generated

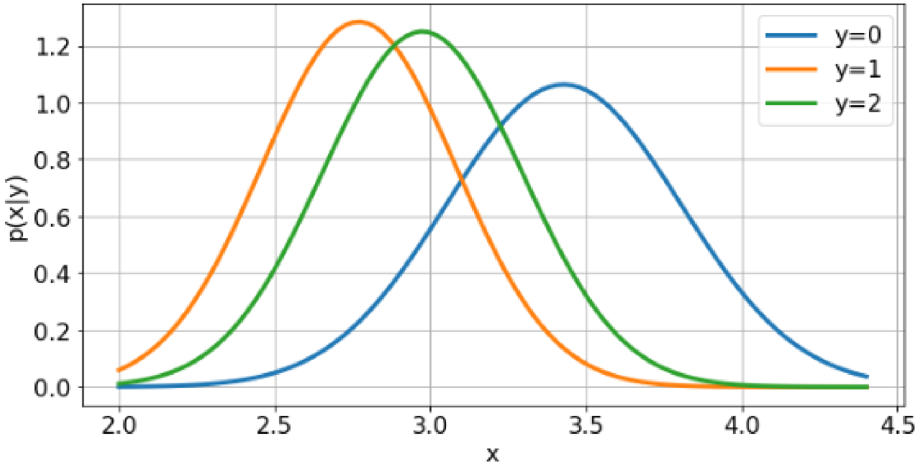
  Example: y ∈ {0, 1} (salmon vs. sea bass) or (airplane vs. bird, etc.)
  x: length of fish

$$p(x|y) = \frac{1}{\sqrt{2\pi}\sigma_y}\exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$$

- μ_y: mean, σ_y^2: variance

Classification

- General classification problem:
  - Assume each sample belongs to one of K classes
  - Observe data on the sample $\boldsymbol{x}$
  - Want to estimate class label $y = 0,1,\ldots,K-1$
  - E.g. dog/cat, spam/real, …



- Strong assumption needed for: decision theory
  - Given each class label y_i, we know conditional distribution $p(\boldsymbol{x}|y_i)$

  Model of how the data is generated
  - We will discuss how we learn this density later…

Maximum Likelihood (ML) Decision

- Which fish type is more likely to given the observed fish length x?

  If $p(x \mid y = 1) > p(x \mid y = 0)$
  guess sea bass;
  otherwise classify the fish as salmon

  - $p(x|y)$ called the likelihood of x given class y
  - Select class with highest likelihood

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

  - Likelihood ratio test (LRT):

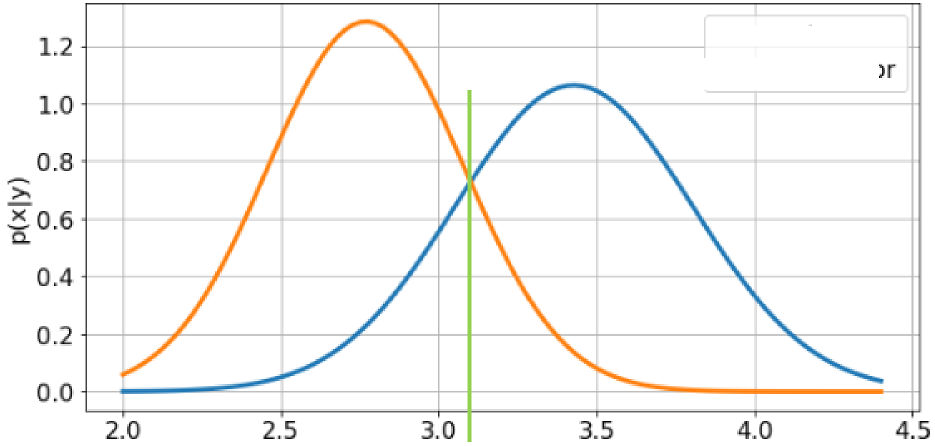
$$\text{If } \frac{p(x \mid y=1)}{p(x \mid y=0)} > 1, \text{ guess sea bass}$$

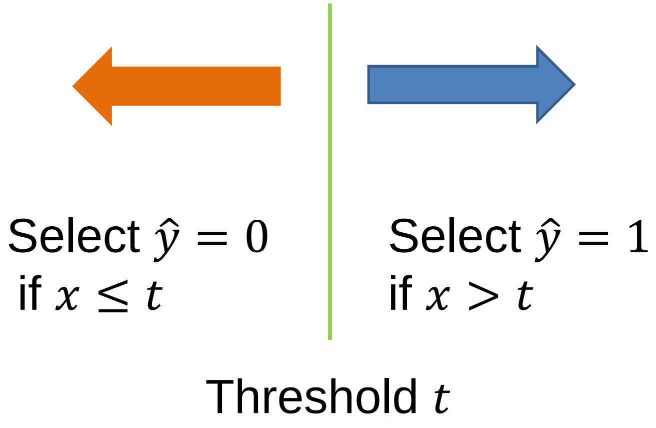
ML Classification

- ML classification: $\hat{y} = \arg\max p(x|y)$
- Binary case: $\hat{y} = \begin{cases} 1 & p(x|1) > p(x|0) \\ 0 & p(x|1) \leq p(x|0) \end{cases}$
- For density on right, we get thresholding decision rule in terms of x:

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

  - t = threshold value where $p(t|1) = p(t|0)$





Likelihood Ratio

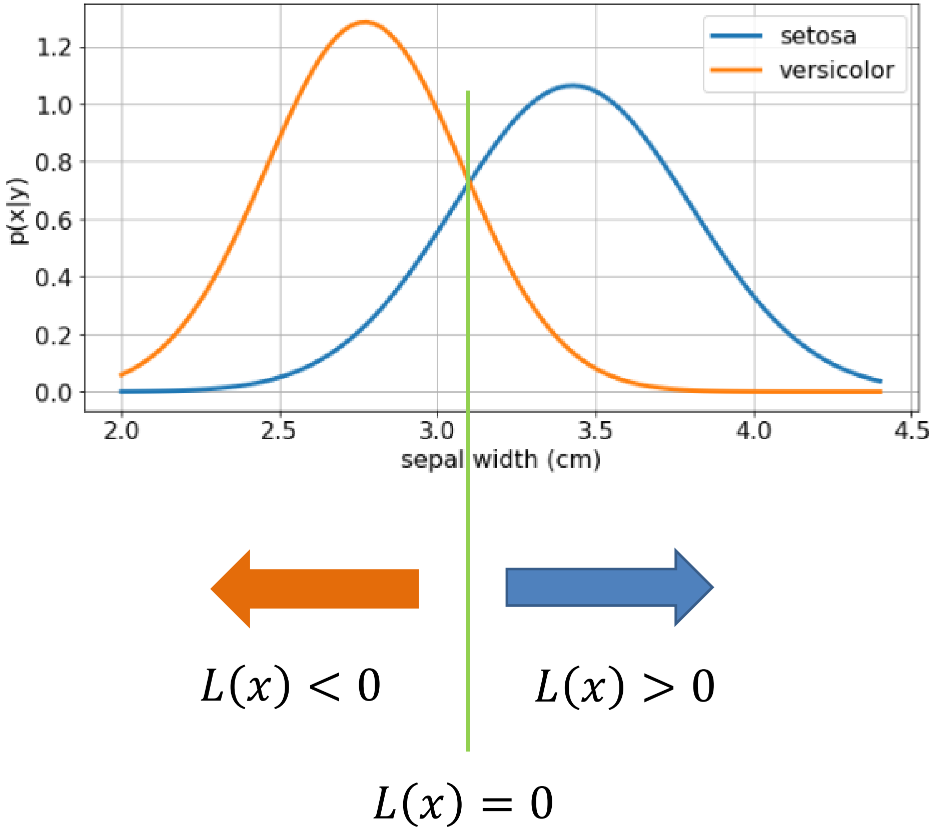
- With likelihoods, it is often easier to work in log domain
- Consider binary classification: $y \in \{0,1\}$
- Define the log likelihood ratio:

$$L(x) \coloneqq \ln \frac{p(x|y=1)}{p(x|y=0)}$$

- ML estimation = likelihood ratio test (LRT):

$$\hat{y} = \begin{cases} 1 & \text{if } L(x) > 0 \\ 0 & \text{if } L(x) \leq 0 \end{cases}$$

- What do we do at boundary?
  - When $L(x) = 0$, we can select either class.
  - Flip a coin, select $y = 0$, select $y = 1$, …
  - It doesn't really matter
  - If x is continuous, probability that $L(x) = 0$ exactly is zero

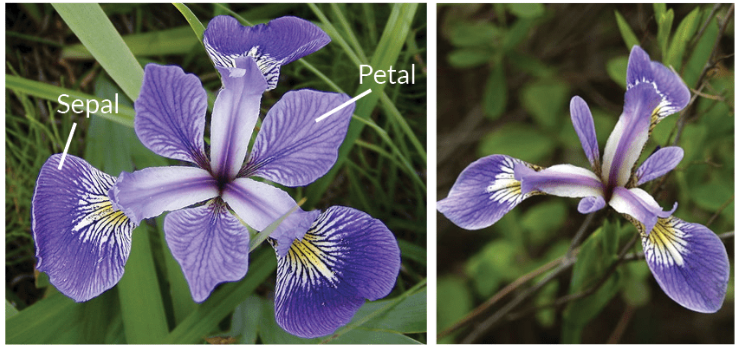


Example: Iris Classification

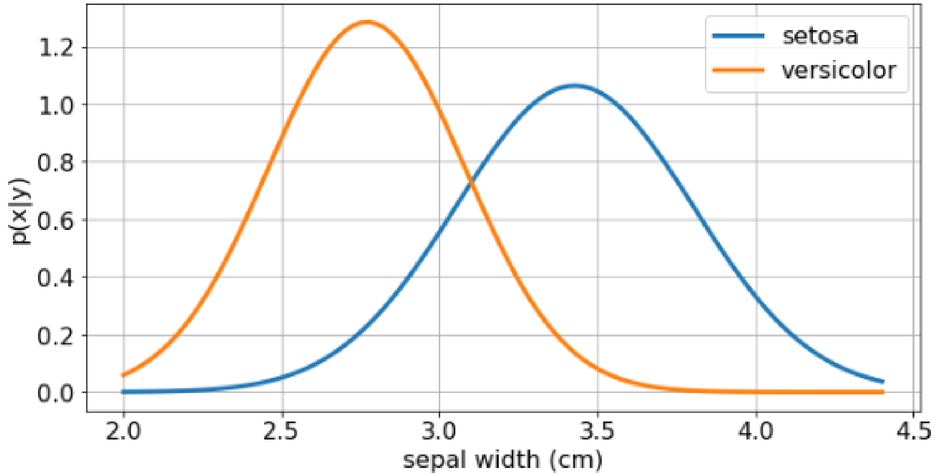
Iris Versicolor | Iris Setosa | Iris Virginica

- Classic Iris dataset used for teaching machine learning
- Get data $\boldsymbol{x} = [x_1, x_2, x_3, x_4]$ for 4 features
  - Sepal length, sepal width, petal length, petal width
  - 150 samples total, 50 samples from each class
- Class label $y \in \{0,1,2\}$ for versicolor, setosa, virginica
- Problem: Learn a classifier for the type of Iris (y) from data $\boldsymbol{x}$

Example: Decision Theory for Iris Classification



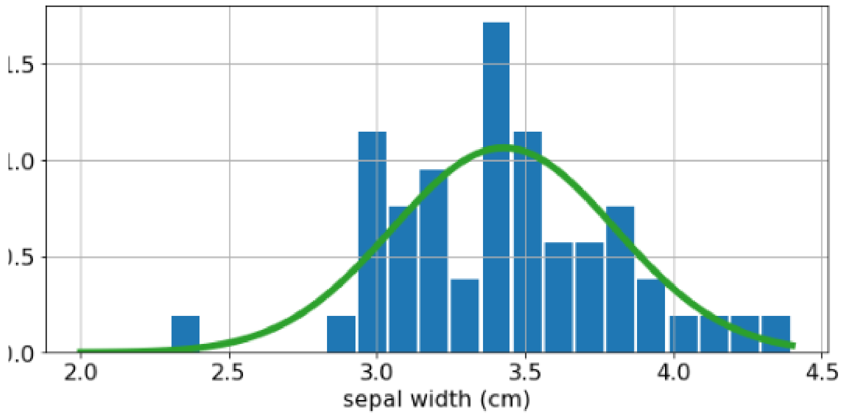
Iris Versicolor **Iris Setosa**

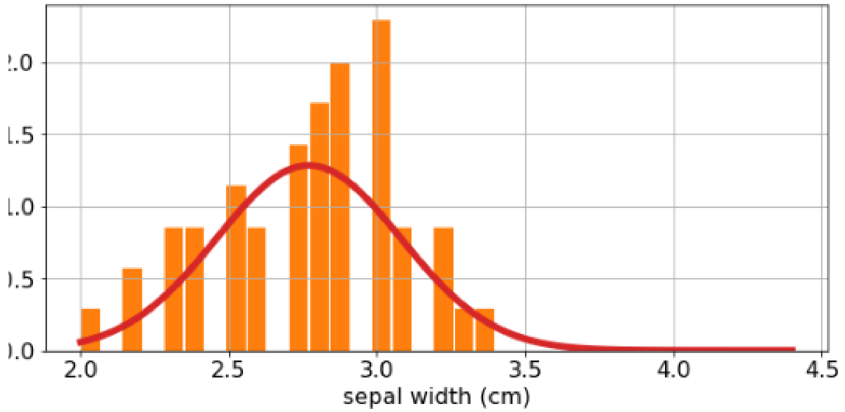


- To make this example simple, assume for now:
  - We classify using only one feature: $x =$sepal width (cm)
  - Select between two classes: Versicolor $(y = 0)$ and Setosa $(y = 1)$
- Also, assume we are given two densities:
  - $p(x|y = 0)$ and $p(x|y = 1)$
  - We assume they are conditionally Gaussian: $p(x|y = k) = N\left(x\middle|\mu_k, \sigma_k^2\right)$
  - Densities represent the condition density of sepal width given the class
  - We will talk about how we get these densities from data later…

How do we get $p(x|y)$?

- Decision theory requires we know $p(\boldsymbol{x}|y)$
  - This is a big assumption!
  - $p(\boldsymbol{x}|y)$ is called the population likelihood
  - Describes theoretical distribution of all samples
- But, in most real problems:
  we have only data samples $(\boldsymbol{x}_i, y_i)$
  - Ex: Iris dataset, we have 50 samples / class
- To use decision theory, we could estimate a density $p(\boldsymbol{x}|y = k)$ for each k from samples
  - Ex: Could assume $p(\boldsymbol{x}|y)$ is Gaussian
  - Estimate mean and variance from samples
- Later, we will talk about:
  - How to do density estimation
  - And if density estimation + decision theory is good idea

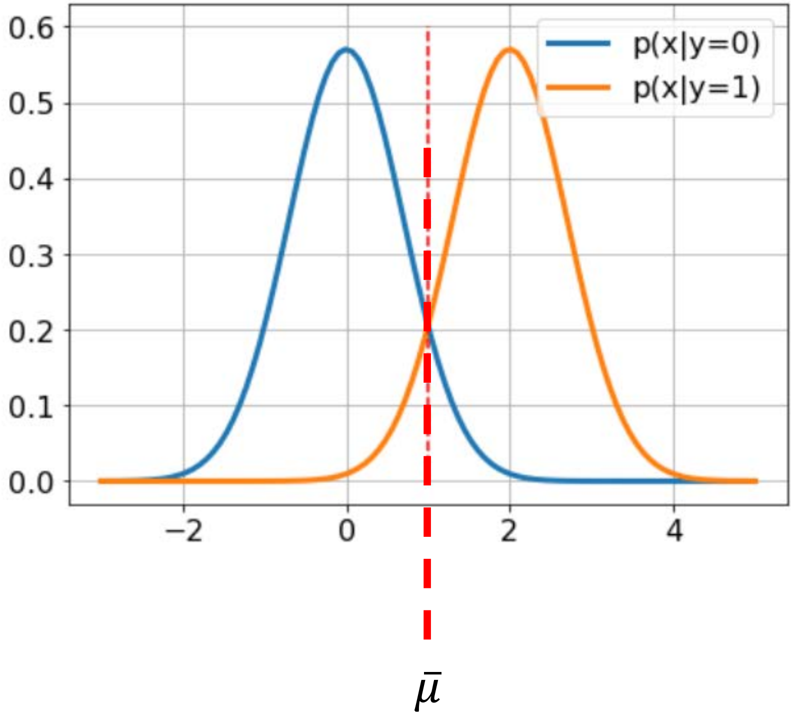




Histograms for two Iris classes
Also plotted is Gaussian with same mean and variance

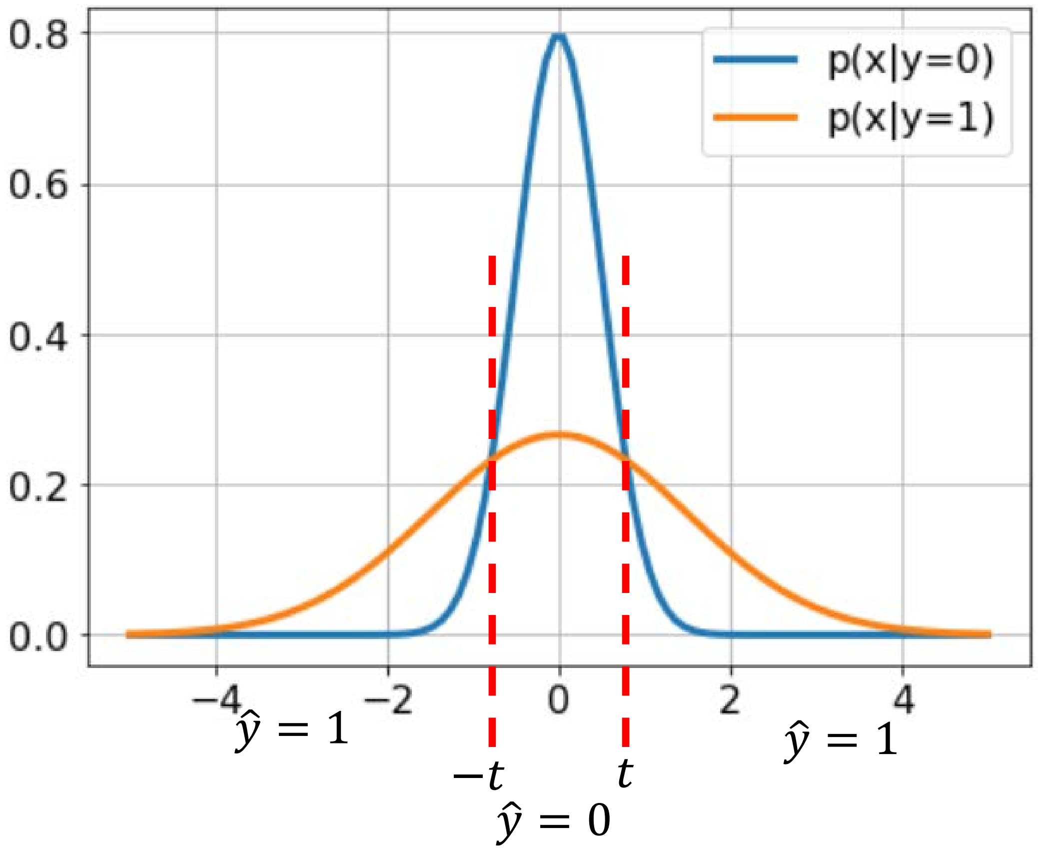
Example Problem: ML for Two Gaussians, Different Means

- Consider binary classification: $y = 0,1$
  - $p(x|y=j) = N(x|\mu_j, \sigma^2), \mu_1 > \mu_0$
  - Two Gaussians with same variance
- Likelihood:
  - $p(x|y=j) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu_i)^2)$
  - $L(x) := \ln \frac{p(x|1)}{p(x|0)} = -\frac{1}{2\sigma^2}[(x-\mu_1)^2 - (x-\mu_0)^2]$
  - With some algebra: $L(x) = \frac{(\mu_1 - \mu_0)}{\sigma^2}[x - \bar{\mu}], \bar{\mu} = \frac{\mu_0 + \mu_1}{2}$
- ML estimate:
  - $\hat{y} = 1 \Leftrightarrow L(x) \geq 0 \Leftrightarrow x \geq \bar{\mu}$
  - With some algebra we get: $\hat{y} = \begin{cases} 1 & \text{if } x > \bar{\mu} \\ 0 & \text{if } x \leq \bar{\mu} \end{cases},$



Example 2: ML for Two Gaussians, Different Variances

- Consider binary classification: $y = 0,1$
  - $p(x|y=j) = N\left(x \middle| 0, \sigma_j^2\right),\ \sigma_0 < \sigma_1$
  - Two Gaussians with different variances, zero mean
- Log likelihood ratio:
  - $p(x|y=j) = \frac{1}{\sqrt{2\pi}\sigma_j} \exp(-\frac{x^2}{2\sigma_j^2})$
  - $L(x) := \ln \frac{p(x|1)}{p(x|0)} = \frac{x^2}{2\sigma_0^2} - \frac{x^2}{2\sigma_1^2} + \frac{1}{2} \ln \frac{\sigma_1^2}{\sigma_0^2}$
- ML estimate:
  - $\hat{y} = 1 \Leftrightarrow L(x) \geq 0 \Leftrightarrow |x| > t$
  - Threshold is $t^2 = \left[\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right]^{-1} \ln \frac{\sigma_1^2}{\sigma_0^2}$



Outline

- Decision Theory
  - Classification, Maximum Likelihood and Log likelihood
  - MAP Estimation, Bayes Risk
  - Probability of errors, ROC
- Empirical Risk Minimization
  - Problems with decision theory, empirical risk minimization
  - Probably approximately correct learning
- Curse of Dimensionality
- Parameter Estimation
  - Probabilistic models for supervised and unsupervised learning
  - ML and MAP estimation
  - Examples

MAP classification

- What if one item is more likely than the other?
- Introduce prior probabilities $P(y=0)$ and $P(y=1)$
  - Salmon more likely than Sea bass: $P(y=0) > P(y=1)$
- Bayes' Rule: $p(y|x) = \frac{p(x \mid y)p(y)}{p(x)}$

- Interested then in class with highest posterior probability $p(y|x)$
- Including prior probabilities:

  If $p(y=0 \mid x) > p(y=1 \mid x)$, guess salmon; otherwise, pick sea bass

- We can write $p(y=0 \mid x) = \frac{p(x|y=0)P(y=0)}{P(x)}$, $p(y=1 \mid x) = \frac{p(x|y=1)P(y=1)}{P(x)}$

MAP classification

- Including prior probabilities:

  If $p(y = 0 \mid x) > p(y = 1 \mid x)$, guess salmon; otherwise, pick sea bass

Maximum A Posterori (MAP) Estimation:

$$\hat{y}_{\text{MAP}} = \alpha(x) = \arg\max_y p(y|x) = \arg\max_y p(x|y)\,P(y)$$

- Select class with highest posterior probability $p(y|x)$

- Binary case: Select $\hat{y}_{\text{MAP}} = 1$ if $p(y = 1|x) > p(y = 0|x)$

From Bayes

$$p(y = 0 \mid x) = \frac{p(x|y=0)P(y=0)}{P(x)},\quad p(y = 1 \mid x) = \frac{p(x|y=1)P(y=1)}{P(x)}$$

Wo we select class 1 if $\dfrac{p(x|y=1)}{p(x|y=0)}\dfrac{P(y=1)}{P(y=0)} \geq 1$

MAP Estimation via LRT

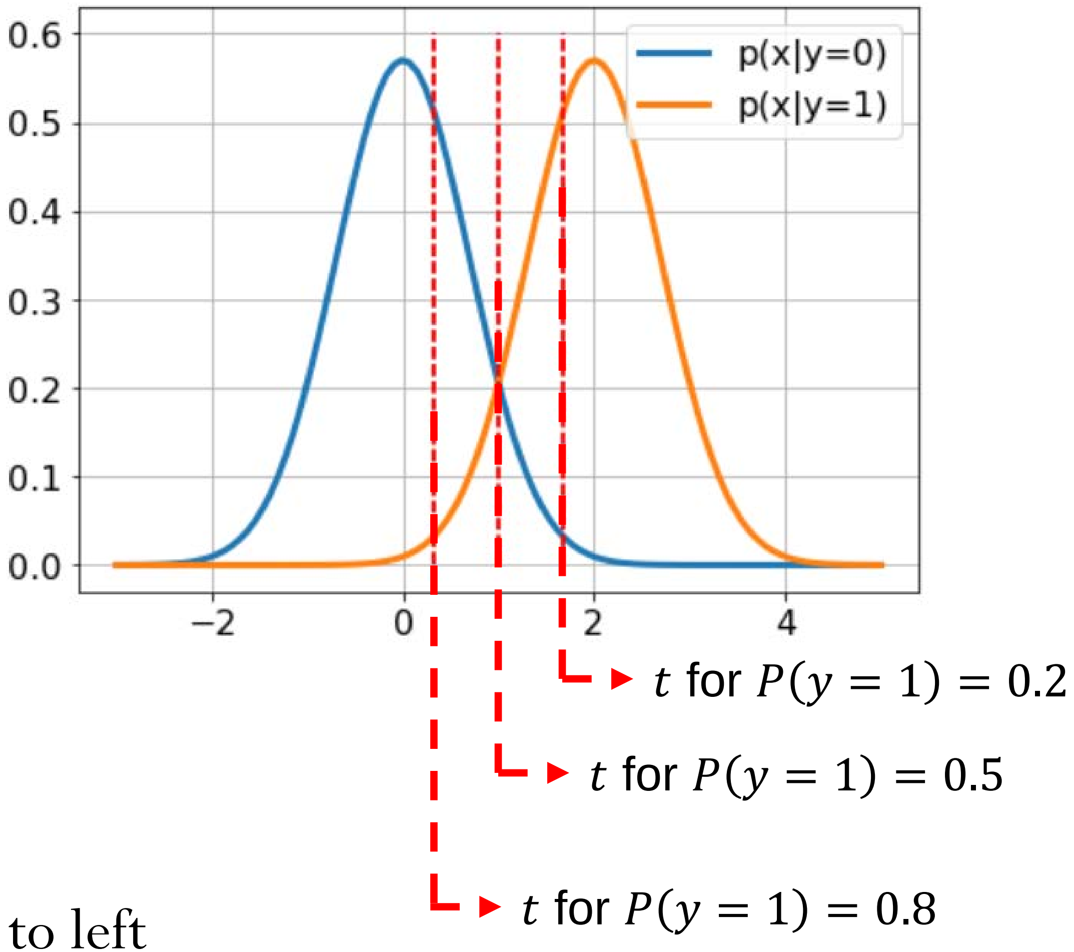
- Consider binary case: $y \in \{0,1\}$
- MAP estimate: Select $\hat{y} = 1 \Leftrightarrow \frac{p(x|y=1)}{p(x|y=0)} \frac{P(y=1)}{P(y=0)} \geq 1 \Leftrightarrow \frac{p(x|y=1)}{p(x|y=0)} \geq \frac{P(y=0)}{P(y=1)}$
- Log domain: select $\hat{y} = 1$ when:

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

  - $L(x) = \ln \left[\frac{p(x|y=1)}{p(x|y=0)}\right]$ is the log likelihood ratio
  - $\gamma = \ln \frac{P(y=0)}{P(y=1)}$ is the threshold for the likelihood function
- In special case where $P(y=1) = P(y=0) = \frac{1}{2}$
  - Threshold is $\gamma = 0$ and MAP estimate becomes identical to ML estimate
- Note you solve this to get it in terms of threshold for x that we denote t

Example: MAP for Two Gaussians, Different Means

- Consider binary classification: $y = 0{,}1$
  - $p(x|y=j) = N(x|\mu_j, \sigma^2), \mu_1 > \mu_2$
  - $P_j = P(y=j)$
- LLRTis:
  - $L(x) = \ln \frac{p(x|1)}{p(x|0)} = \frac{(\mu_1 - \mu_0)(x - \bar{\mu})}{\sigma^2} \quad \bar{\mu} = \frac{\mu_0 + \mu_1}{2}$
- MAP estimate: Let $\gamma = \ln \frac{P_0}{P_1}$
  - $\hat{y} = 1 \Leftrightarrow L(x) \geq \gamma \Leftrightarrow x \geq \bar{\mu} + \frac{\sigma^2 \gamma}{\mu_1 - \mu_0}$
  - Threshold is shifted by the prior probability γ
  - If $P(y=1) > P(y=0) \Rightarrow \gamma < 0 \Rightarrow t$ is shifted to left $\Rightarrow$ Estimator more likely to select $\hat{y} = 1$



Often more formally written Hypothesis Testing

- Two possible hypotheses for data
  - H_0: Null hypothesis, $y = 0$
  - H_1: Alternate hypothesis, $y = 1$
- Model statistically:
  - $p(x|H_i), i = 0,1$
  - Assume some distribution for each hypothesis
- Given
  - Likelihood $p(x|H_i), i = 0,1$, Prior probabilities $p_i = P(H_i)$
- Compute posterior $P(H_i|x)$
  - How likely is H_i given the data and prior knowledge?
- Bayes' Rule:

$$P(H_i|x) = \frac{p(x|H_i)p_i}{p(x)} = \frac{p(x|H_i)p_i}{p(x|H_0)p_0 + p(x|H_1)p_1}$$

MAP: Minimum Probability of Error

- Probability of error:

$$P_{err} = P(\hat{H} \neq H) = P(\hat{H} = 0 | H_1) p_1 + P(\hat{H} = 1 | H_0) p_0$$

- Write with integral:

$$P(\hat{H} \neq H) = \int p(x) P(\hat{H} \neq H | x) dx$$

- It can be shown (you won't have to) that error is minimized with MAP estimator

$$\hat{H} = 1 \Leftrightarrow P(H_1|x) \geq P(H_0|x)$$

- Key takeaway: MAP estimator minimizes the probability of error

Making it more interesting, full on Bayes

- What does it cost for a mistake? Plane with a missile, not a big bird?
- Define loss or cost:

  $L(\alpha(x), y)$: cost of decision $\alpha(x)$ when state is y

  also often denoted C_{ij}

	Y = 0	Y = 1
$\alpha(x) = 0$	Correct, cost L(0,0)	Incorrect, cost L(0,1)
$\alpha(x) = 1$	incorrect, cost L(1,0)	Correct, cost L(1,1)

- Classic: Pascal's wager

	God exists (G)	**God does not exist (¬G)**
Belief (B)	+∞ (infinite gain)	−1 (finite loss)
Disbelief (¬B)	−∞ (infinite loss)	+1 (finite gain)

Risk Minimization

- So now we have: the likelihood functions p(x | y)

  priors p(y)

  decision rule $\alpha(x)$

  loss function $L(\alpha(x), y)$:
- *Risk* is expected loss:

$$
\begin{aligned}
E[L] = & \cancel{L(0,0)}\, p(\alpha(x) = 0, y = 0) \\
& + L(0,1)\, p(\alpha(x) = 0, y = 1) \\
& + L(1,0)\, p(\alpha(x) = 1, y = 0) \\
& + \cancel{L(1,1)}\, p(\alpha(x) = 1, y = 1)
\end{aligned}
$$

- Without loss of generality, zero cost for correct decisions

$$
\begin{aligned}
E[L] = & \ L(1,0)\, p(\alpha(x) = 1 \mid y = 0) p(y = 0) \\
& + L(0,1)\, p(\alpha(x) = 0 \mid y = 1) p(y = 1)
\end{aligned}
$$

- Bayes Decision Theory says "pick decision rule $\alpha(x)$ to minimize risk"

Bayes Risk Minimization

- As before, express risk as integration over x:

$$R = \int \sum_{ij} C_{ij}\, P(y = j|x) 1_{\{\hat{y}(x)=i\}}\, p(x) dx$$

- To minimize, select $\hat{y} = 1$ when
  - $C_{10}P(y = 0|x) + C_{11}P(y = 1|x) \leq C_{00}P(y = 0|x) + C_{01}P(y = 1|x)$
  - $P(y = 0|x)/P(y = 1|x) \geq (C_{10} - C_{00})/(C_{11} - C_{01})$

- By Bayes Theorem, equivalent to an LRT with

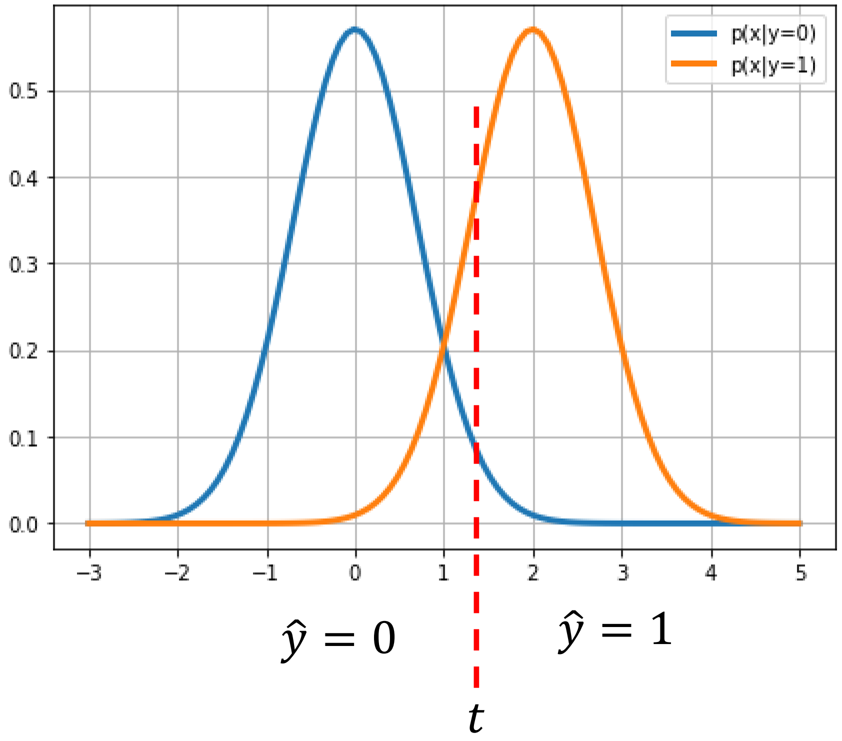
$$\frac{P(x|y = 1)}{P(x|y = 0)} \geq \frac{(C_{10} - C_{00})p_0}{(C_{11} - C_{01})p_1}$$

Outline

- Decision Theory
  - Classification, Maximum Likelihood and Log likelihood
  - MAP Estimation, Bayes Risk
  - Probability of errors, ROC
- Empirical Risk Minimization
  - Problems with decision theory, empirical risk minimization
  - Probably approximately correct learning
- Curse of Dimensionality
- Parameter Estimation
  - Probabilistic models for supervised and unsupervised learning
  - ML and MAP estimation
  - Examples

Computing Error Probabilities

- How do we compute errors?
- Suppose that decision rule is of the form: $\hat{y} = \begin{cases} 1 & \text{if } g(x) > t \\ 0 & \text{if } g(x) \le t \end{cases}$
  - $g(x)$ is called the discriminator
  - t is the threshold
- Ex: Decision rule for scalar Gaussians
  - $\hat{y} = \begin{cases} 1 & \text{if } x > t \\ 0 & \text{if } x \le t \end{cases}$
  - Uses a linear discriminator $g(x) = x$
  - Threshold t will depend on estimator type
    ML, MAP, Bayes risk, ..
- We will compute the error as a function of t



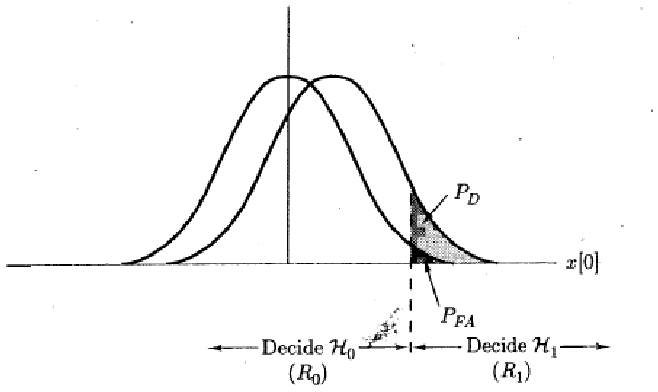
Types of Errors

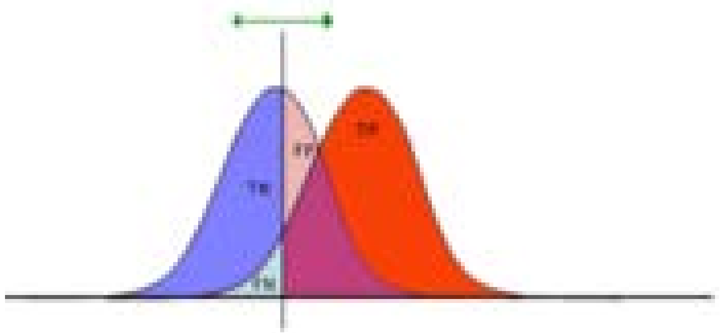
- Consider binary case: $y \in \{0,1\}$
- Two possible errors:
  - Type I error (False alarm or False Positive): Decide $\hat{y} = 1$ when $y = 0$
  - Type II error (Missed detection or False Negative): Decide $\hat{y} = 0$ when $y = 1$

- The effect of the errors may be very different
- Example: Disease diagnosis: $y = 1$ patient has disease, $y = 0$ patient is healthy
  - Type I error: You say patient is sick when patient is healthy
    Error can cause extra unnecessary tests, stress to patient, etc…
  - Type II error: You say patient is fine when patient is sick
    Error can miss the disease, disease could progress, …

Visualizing Errors

- Type I error (False alarm or False Positive): Decide H1 when H0
- Type II error (Missed detection or False Negative): Decide H0 when H1
- Trade off
- Can work out error probabilities from conditional probabilities

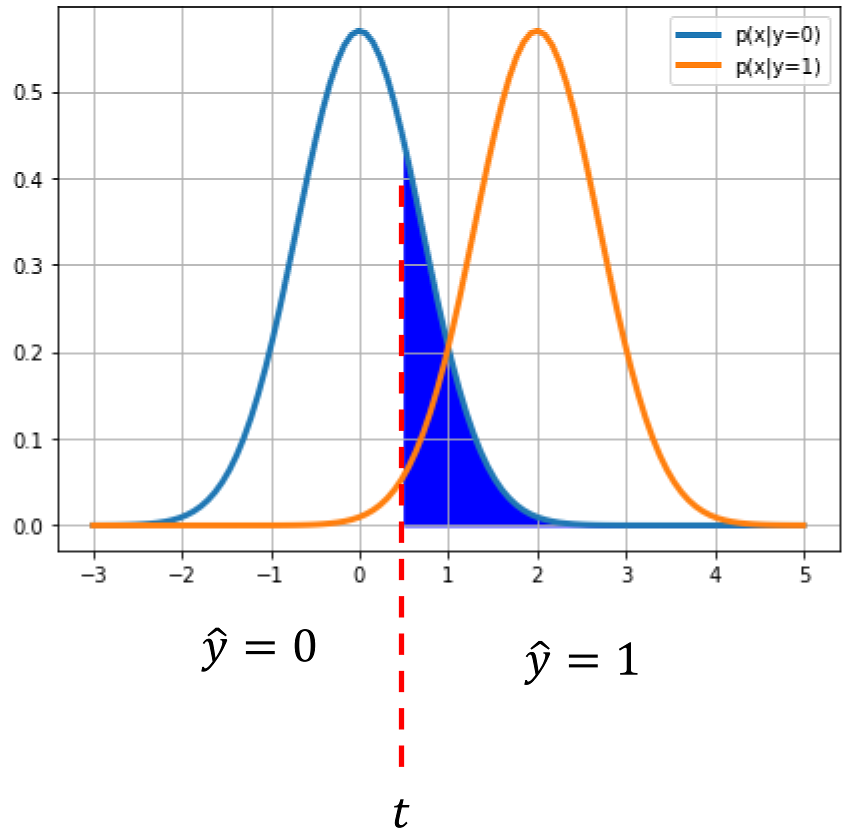




TP	FP
FN	TN
1	1

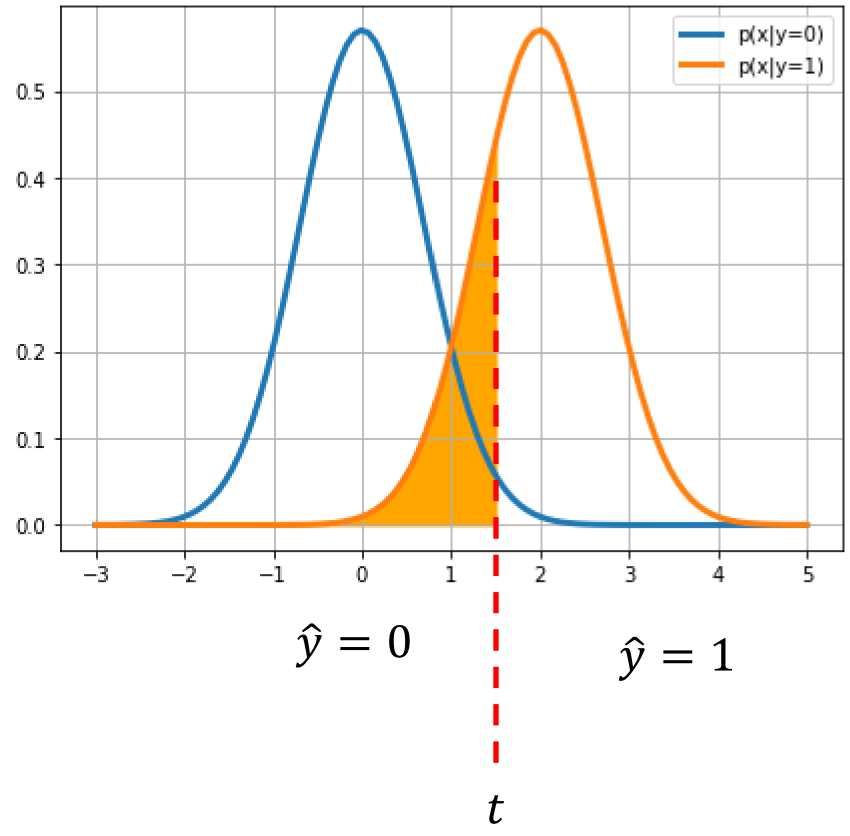
Scalar Gaussian Example: False Alarm

- Scalar Gaussian: For $j = 0{,}1$:
  - $p(x|y=j) = N(x|\mu_j, \sigma^2), \mu_1 > \mu_0$
- False alarm:
  - $P_{FA} = P(\hat{y}=1|y=0) = P(x \geq t|y=0)$
  - This is the area under curve, $P_{FA} = \int_t^\infty p(x|y=0)\,dx$
  - But, we can compute this using Gaussians
  - Given $y=0,\ x \sim N(\mu_0, \sigma^2)$
  - Therefore: $P_{FA} = P(x \geq t|y=0) = Q\left(\frac{t-\mu_0}{\sigma}\right)$



Scalar Gaussian Example: Missed Detection

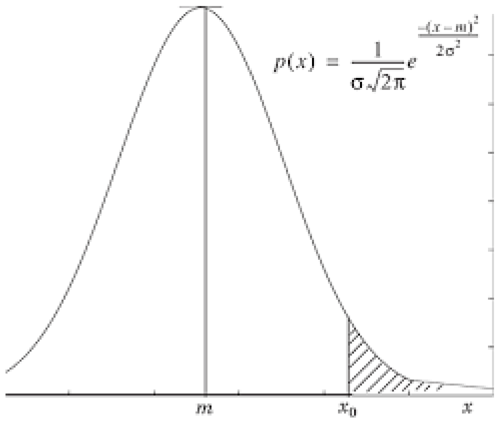
- Scalar Gaussian: For $j = 0,1$:
  - $p(x|y = j) = N(x|\mu_j, \sigma^2), \mu_1 > \mu_0$
- Missed detection can be computed similarly
  - $P_{MD} = P(\hat{y} = 0|y = 1) = P(x \leq t|y = 1)$
  - This is the area under curve
  - But, we can compute this using Gaussians
  - Given $y = 1,\ x \sim N(\mu_1, \sigma^2)$
  - Therefore: $P_{FA} = P(x \leq t|y = 1) = 1 - Q\left(\frac{t-\mu_1}{\sigma}\right)$



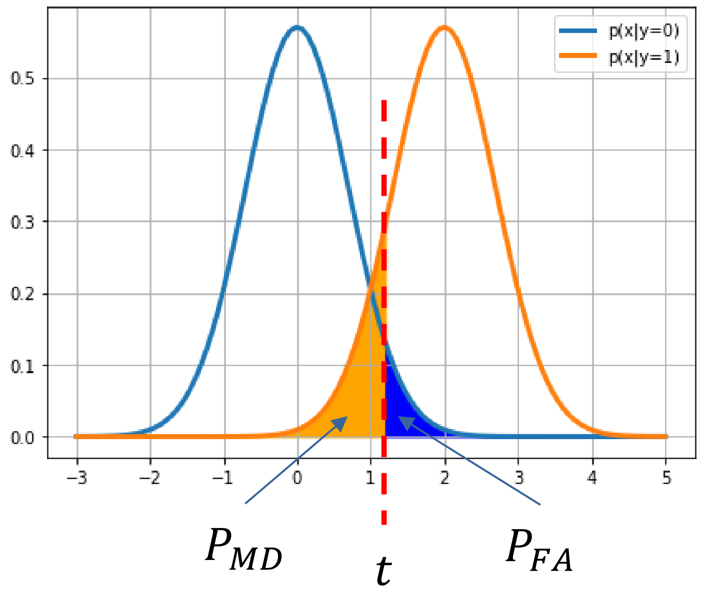
Review: Gaussian Q-Function

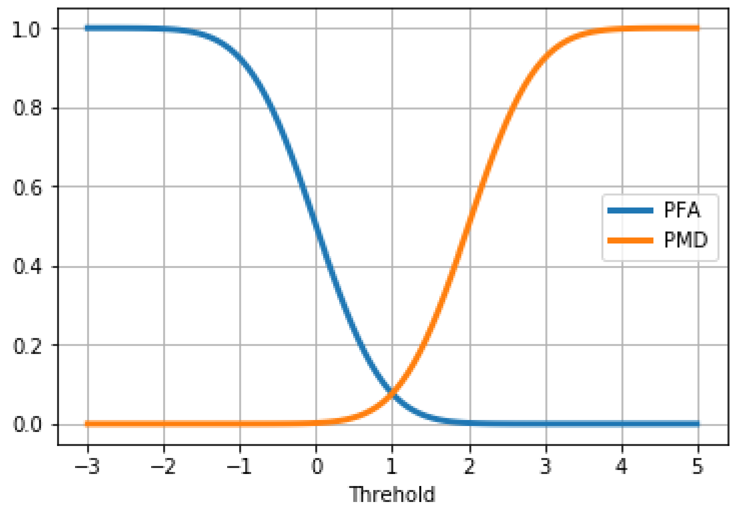
- Problem: Suppose $X \sim N(\mu, \sigma^2)$.
  - Often must compute probabilities like $P(X \geq t)$
  - No closed-form expression.
- Define Marcum Q-function:
  $Q(z) = P(Z \geq z),\ Z \sim N(0,1)$
- Let $Z = (X - \mu)/\sigma$
- Then

$$P(X \geq t) = P\left(Z \geq \frac{t-\mu}{\sigma}\right) = Q\left(\frac{t-\mu}{\sigma}\right)$$



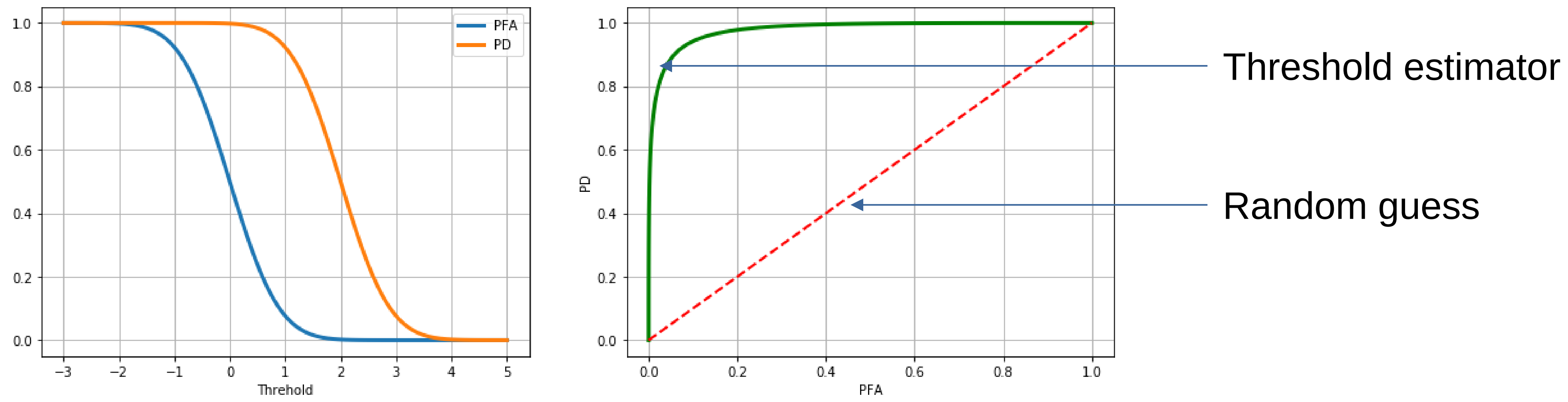
FA vs. MD Tradeoff





- We see that there is a tradeoff:
  - Increasing threshold $t \Rightarrow$ Decreases P_{FA}
  - But, increasing threshold $t \Rightarrow$ Increases P_{MD}
- What threshold value we select depends on their relative costs
  - What is the effect of a FA vs. MD
  - Consider medical diagnosis case

ROC Curve



- Receiver Operating Characteristic (ROC) curve
  - For each threshold level t compute $P_D(t) = 1 - P_{MD}(t)$ and $P_{FA}(t)$
  - Plot $P_D(t)$ vs. $P_{FA}(t)$
  - Shows the how large the detection probability can be for a given P_{FA}
  - Name "ROC" comes from communications receivers where these were first used
- Comparing ROC curves
  - Higher curve is better
  - Random guessing gets red line: Guess $\hat{y} = 1$ with probability t
  - So, any decent estimator should be above the red line

Multiple Classes

- Often have multiple classes. $y \in 1, \ldots, K$
- Most methods easily extend:
  - ML: Take max of K likelihoods:

$$\hat{y} = \arg \max_{i=1,\ldots,K} p(x|y=i)$$

  - MAP: Take max of K posteriors:

$$\hat{y} = \arg \max_{i=1,\ldots,K} p(y=i|x) = \arg \max_{i=1,\ldots,K} p(x|y=i)p(y=i)$$

  - LRT: Take max of K weighted likelihoods:

$$\hat{y} = \arg \max_{i=1,\ldots,K} p(x|y=i)\,\gamma_i$$

Outline

- Decision Theory
  - Classification, Maximum Likelihood and Log likelihood
  - MAP Estimation, Bayes Risk
  - Probability of errors, ROC
- Empirical Risk Minimization
  - Problems with decision theory, empirical risk minimization
  - Probably approximately correct learning
- Curse of Dimensionality
- Parameter Estimation
  - Probabilistic models for supervised and unsupervised learning
  - ML and MAP estimation
  - Examples

Two Approaches

- Bayesian formulation for classification: Requires we know $p(x|y)$
- But, we only have samples $(x_i, y_i), i = 1, \ldots, N$, from this density
- What do we do?

- Approach 1: Probabilistic approach
  - Learn distributions $p(x|y)$ from data (x_i, y_i)
  - Then apply Bayesian decision theory using estimated densities

- Approach 2: Decision rule
  - Use hypothesis testing to select a form for the classifier
  - Learn parameters of the classifier directly from data

Example with Scalar Data and Linear Discriminator

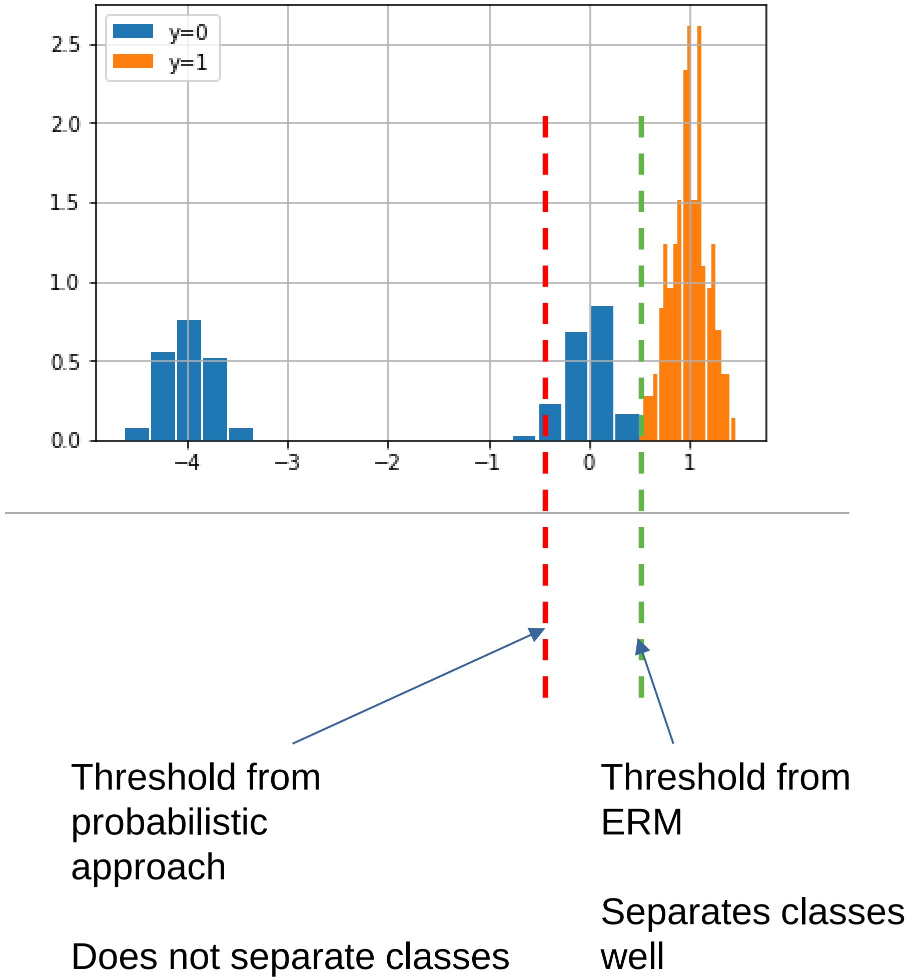
- Given data $(x_i, y_i), i = 1, \dots, N$
- Probabilistic approach:
  - Assume $x_i \sim N(\mu_0, \sigma^2)$ when $y_i = 0$; $x_i \sim N(\mu_1, \sigma^2)$ when $y_i = 1$
  - Learn sample means for two classes: $\hat{\mu}_j =$ mean of samples x_i in class j
  - From decision theory, we have the decision rule:

$$\hat{y} = \alpha(x, t) = \begin{cases} 1 & x > t \\ 0 & x < t \end{cases}, \qquad t = \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}$$

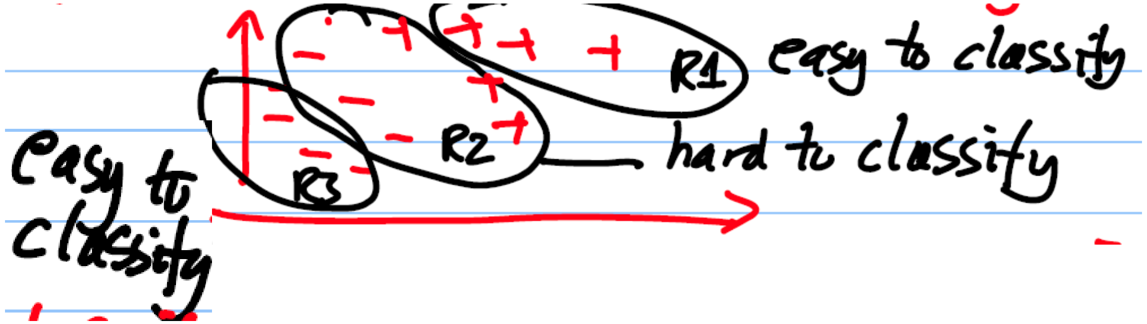
- Empirical Risk minimization
  - For each threshold t, we get decisions on the training data: $\hat{y}_i = \alpha(x_i, t)$
  - Look at empirical risk, e.g. training error $L(t) := \frac{1}{N} \#\{\hat{y}_i \neq y_i\}$
  - Select t to minimize empirical risk $\hat{t} = \arg\min_t L(t)$

Why ERM may be Better

- Suppose data is as shown
- We estimate class means: $\hat{\mu}_0 \approx -2, \quad \hat{\mu}_1 \approx 1$
- Decision rule from probabilistic approach
  - $\hat{y} = \begin{cases} 1 & x > t \\ 0 & x < t \end{cases}, \quad t = \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} \approx -0.5$
  - Threshold misclassifies many points
- Empirical risk minimization
  - Select t to minimize classification errors on training data
  - Will get $t \approx 0.5$ ⇒Leads to better rule
- Why probabilistic approach failed?
  - We assumed both distributions were Gaussian
  - But, $p(x|y = 0)$ is not Gaussian. It is bimodal
  - ERM does not require such assumptions



Example of Decision Rule Approach

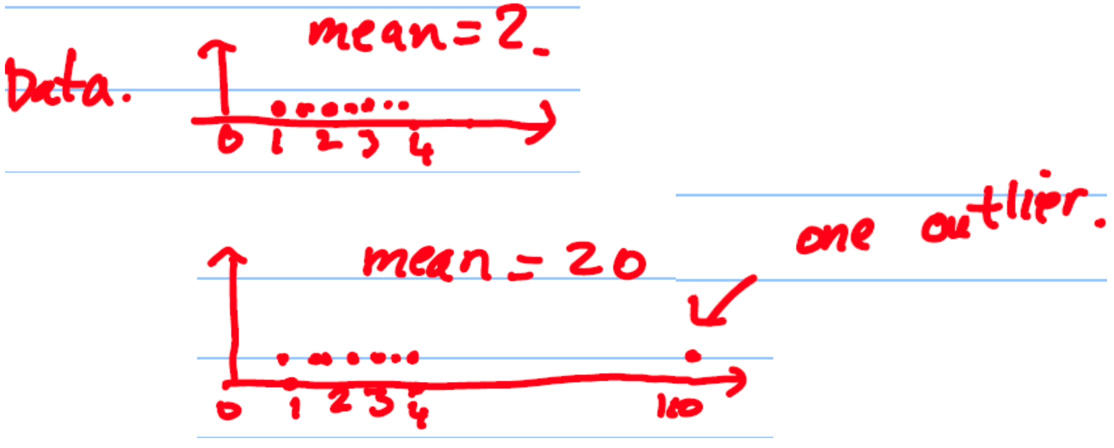


- Decision rule approach :
  - Assume a rule: $\hat{y} = \alpha(x) = \begin{cases} 1 & x > t \\ 0 & x < t \end{cases}$
  - Rule has an unknown parameter t
  - Find t to minimize empirical risk $R_{\text{emp}}(\alpha, X_N) := \frac{1}{N} \sum_i 1(y_i \neq \alpha(x_i))$
  - Minimizes error on training data

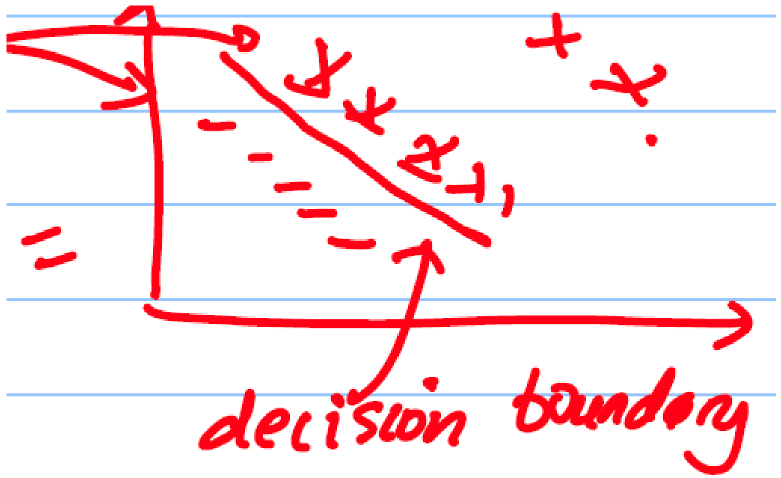
- Motivation for decision rule approach over probabilistic approach
  - Why bother learning probabilities densities if your final goal is a decision rule
  - Assumptions on probability densities may be incorrect (see next slide)
  - Concentrate your efforts by dealing with data that is hard to classify

Dangers of Using Probabilistic Approach

- Needs to assume specific form of densities
- Ex: Suppose we assume Gaussian densities
  - Gaussians are not robust
  - Outlier values can make large changes in mean and variance estimates



- Risk minimization alternative:
  - Search over planes that separates classes
  - Only pay attention to data near boundary
  - Good in case of limited data



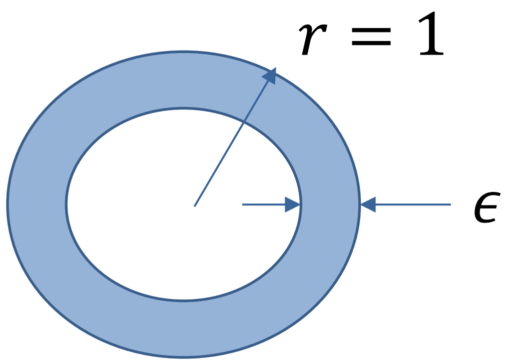
Outline

- Decision Theory
  - Classification, Maximum Likelihood and Log likelihood
  - MAP Estimation, Bayes Risk
  - Probability of errors, ROC
- Empirical Risk Minimization
  - Problems with decision theory, empirical risk minimization
  - Probably approximately correct learning
- Curse of Dimensionality
- Parameter Estimation
  - Probabilistic models for supervised and unsupervised learning
  - ML and MAP estimation
  - Examples

Intuition in High-Dimensions

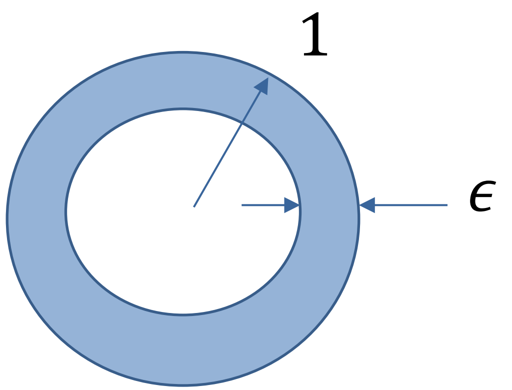
- Examples of Bayes Decision theory can be misleading
  - Examples are in low dimensional spaces, 1 or 2 dim
  - Most machine learning problems today have high dimension
  - Often our geometric intuition in high-dimensions is wrong

- Example: Consider volume of sphere of radius $r = 1$ in D dimensions
  - What is the fraction of volume in a thin shell of a sphere between $1 - \epsilon \leq r \leq 1$?



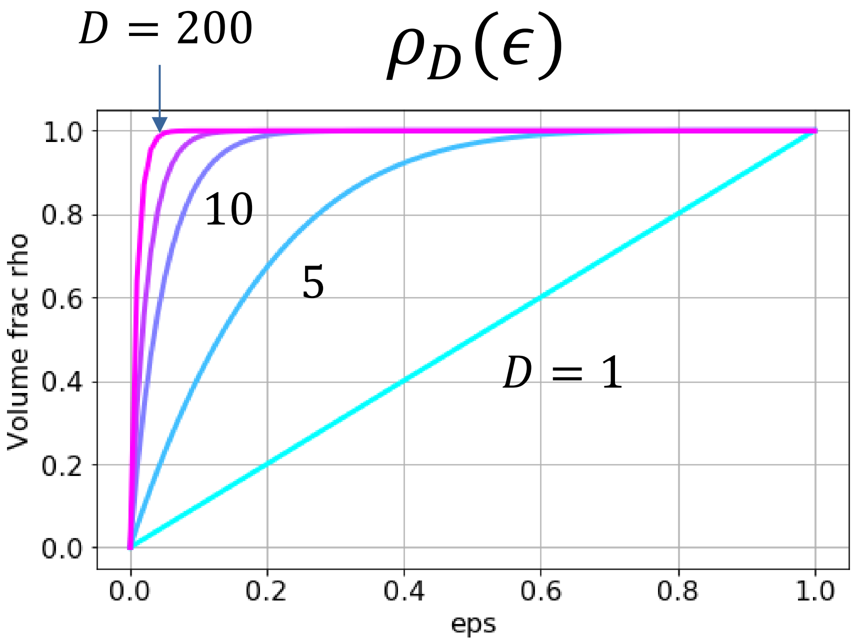
Example: Sphere Hardening

- Let $V_D(r) =$ volume of sphere of radius r, dimension D
  - From geometry: $V_D(r) = K_D r^D$
- Let $\rho_D(\epsilon) =$ fraction of volume in a shell of thickness ϵ



$$\rho_D(\epsilon) = \frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)}$$

$$= \frac{K_D - K_D(1-\epsilon)^D}{K_D} = 1 - (1-\epsilon)^D$$

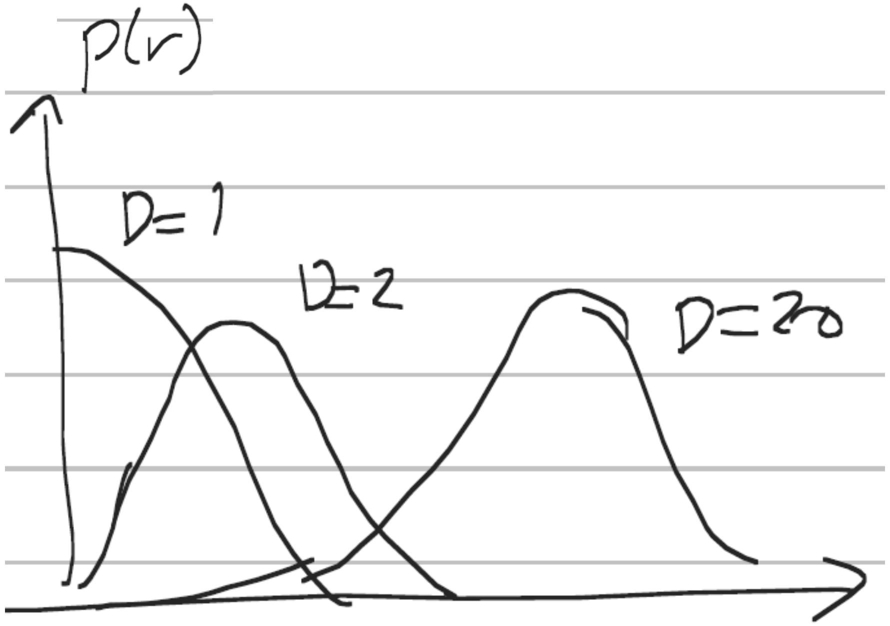


- For any ϵ, we see as $\rho_D(\epsilon) \to 1$ as $D \to \infty$
- All volume concentrates in a thin shell
- This is very different than in low dimensions

Gaussian Sphere Hardening

- Consider a Gaussian i.i.d. vector
  - $x = (x_1, \ldots, x_D), \;\; x_i \sim N(0,1)$
- As $D \to \infty$, probability density concentrates on shell $\|x\| \approx \sqrt[2]{D}$, even though $x = 0$ is most likely point

- Let $r = (x_1^2 + x_2^2 + \cdots + x_D^2)^{1/2}$
  - $D = 1$: $p(r) = c\, e^{-r^2/2}$
  - $D = 2$: $p(r) = c\, r\, e^{-r^2/2}$
  - general D: $p(r) = c\, r^{D-1}\, e^{-r^2/2}$



Example: Sphere Hardening

- Conclusions: As dimension increases,
  - All volume of a sphere concentrates at its surface!

- Similar example: Consider a Gaussian i.i.d. vector
  - $x = (x_1, \dots, x_d), \;\; x_i \sim N(0,1)$
  - As $d \to \infty$, probability density concentrates on shell
$$\|x\|^2 \approx d$$
  - Even though $x = 0$ is most likely point

Computational Issues

- In high dimensions, classifiers need large number of parameters
- Example:
  - Suppose $x = (x_1, \dots, x_d)$, each x_i takes on L values
  - Hence x takes on L^d values
- Consider general classifier $f(x)$
  - Assigns each x some value
  - If there are no restrictions on $f(x)$, needs L^d paramters

Curse of Dimensionality

- Curse of dimensionality: As dimension increases
  - Number parameters for functions grows exponentially
- Most operations become computationally intractable
  - Fitting the function, optimizing, storage

- What ML is doing today
  - Finding tractable approximate approaches for high-dimensions