Classification – 1 (Introduction)

Aside: Decision Trees (KM 18.1, EA Ch 9)

(non-parametric, discriminative)

Fundamentals: Decision Theory (CB 1.5, KM 5-5.1.2)

Evaluation: Measuring Performance of a Classifier (KM 5.1.3, 5.1.4)

What is Classification?

- A predictive modeling technique with categorical outputs (labels)
  - Input: training records, each with a class label
  - Build: a model that can predict the label of future records of unknown class

- **METHODS:**
  - (A) **Discriminative Approaches**: Focus on decision boundary (not on underlying probability models)
    - decision trees (C4.5, CART, CHAID,...)
    - SVM
  - (B) **Generative Approaches**: Focus on class membership probability (statistical/ Pattern recognition)

  This philosophy "recognizes the probabilistic nature both of the information we seek to process, and of the form in which we should express the results".
    - Bayesian, k-nearest neighbor, logistic regression
    - neural networks

A (Hard) Classification Model Partitions The Feature Space

- E.g. grading apples based on size and shine.

- Obtaining decision boundaries
  - Explicit
    - Youth if age < 18
  - vs. Implicit
    - Partitioning via discriminant functions $d_i(x)$, one for each class *i*.
      - Object x belongs to class i iff $d_i(x) > d_j(x)$
      - e.g., $d_i(x) = P(C_i|x)$

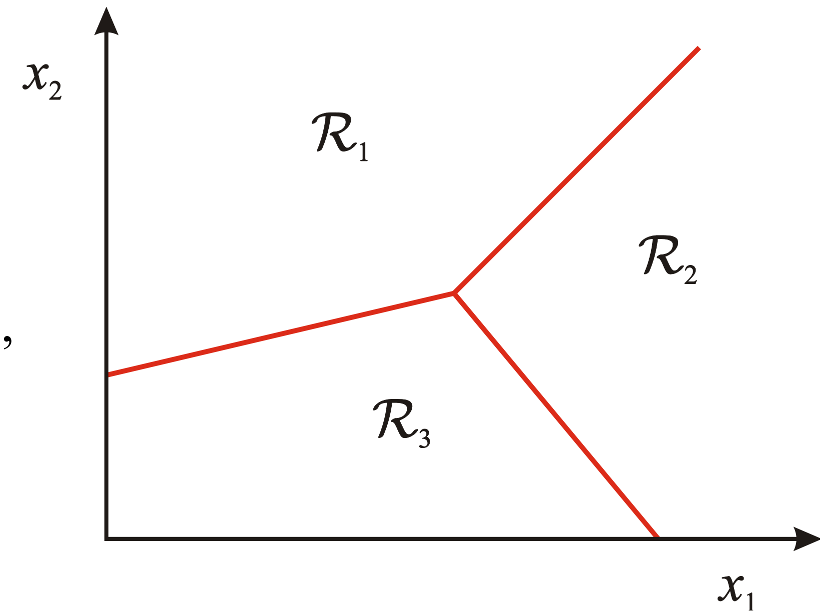


Fig: A 2-D input space partitioned into 3 regions for a 3 class problem

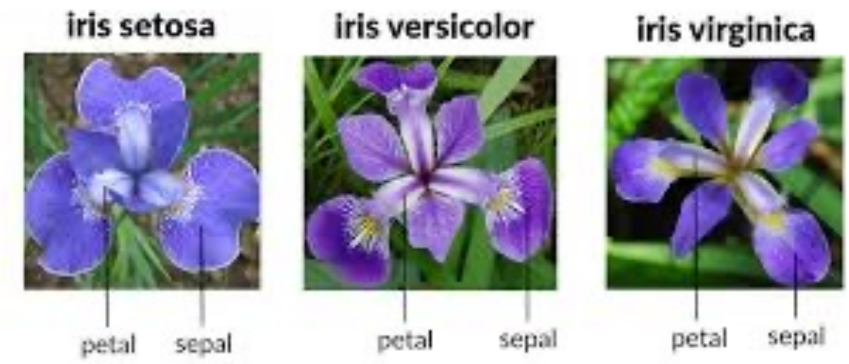
Decision Trees

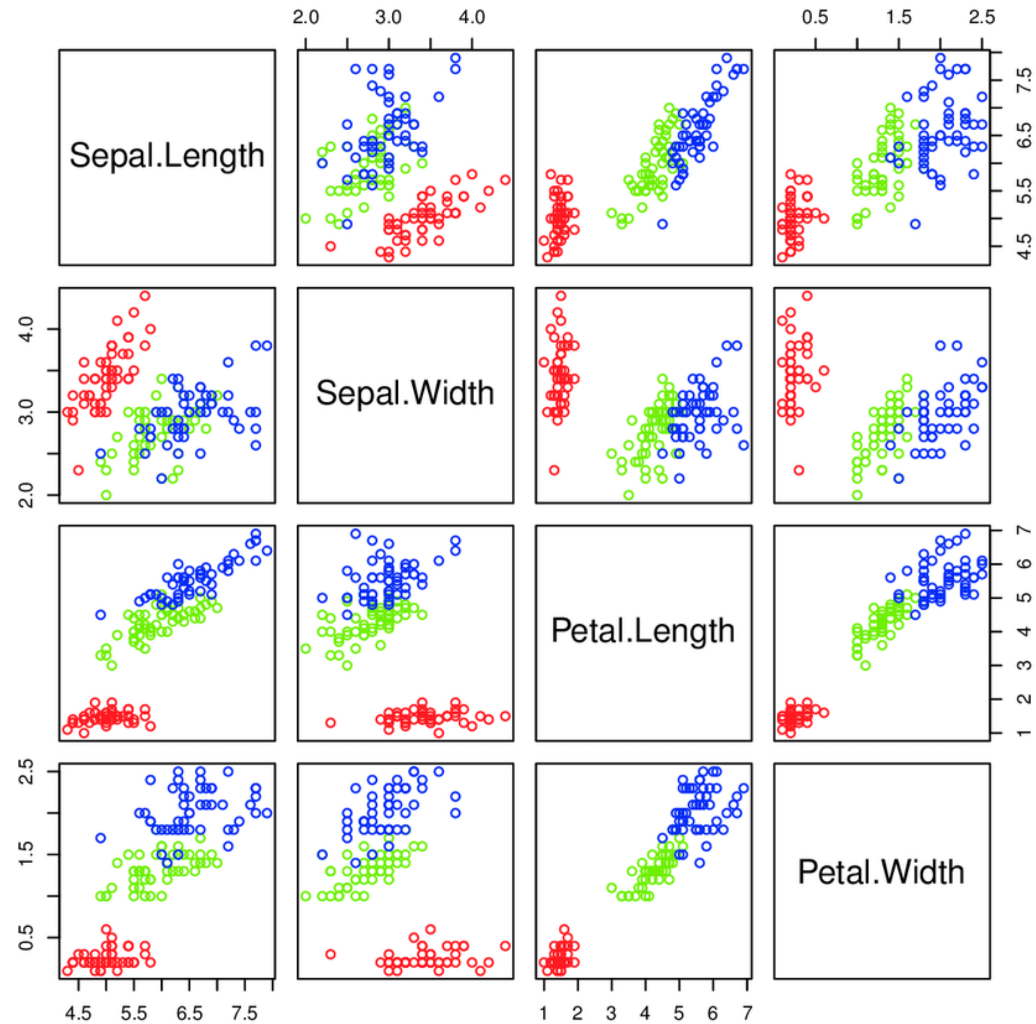
(Simple but Popular Classifiers)

Non-parametric!

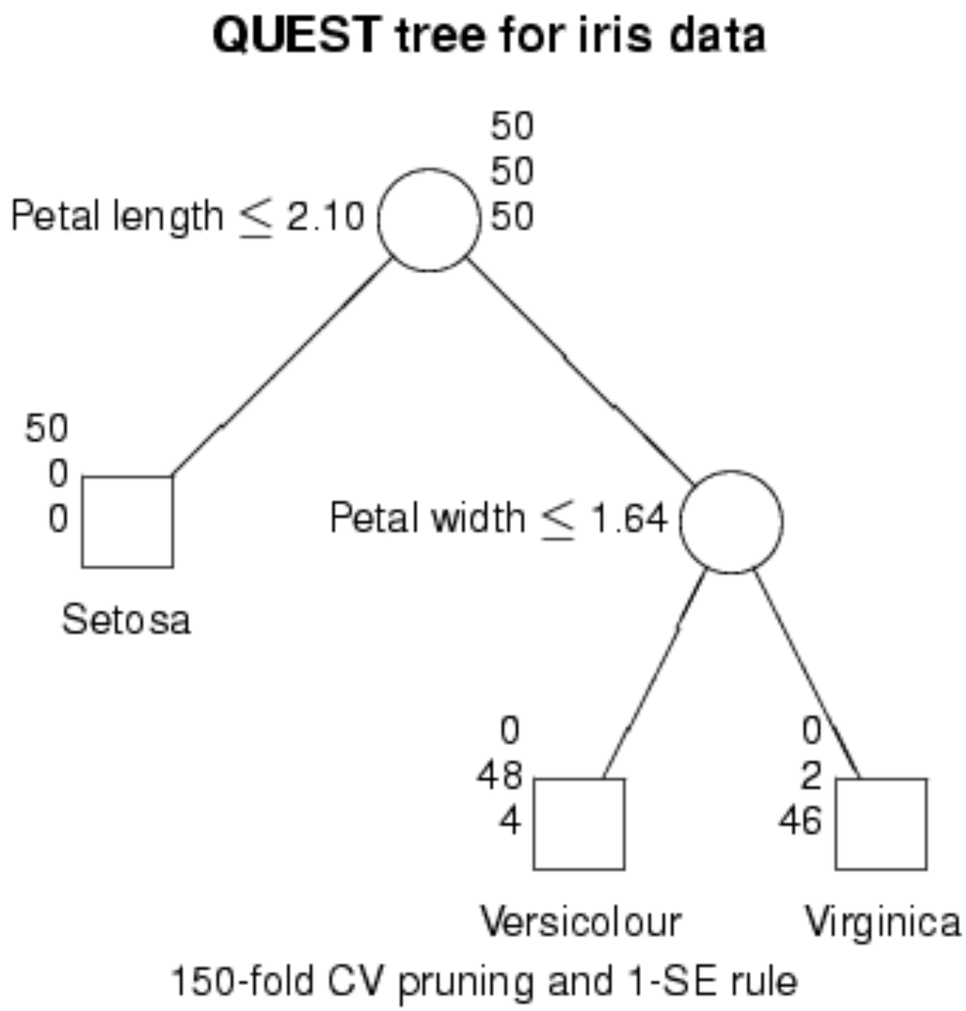
Read: KM 18.1, EA Ch 9

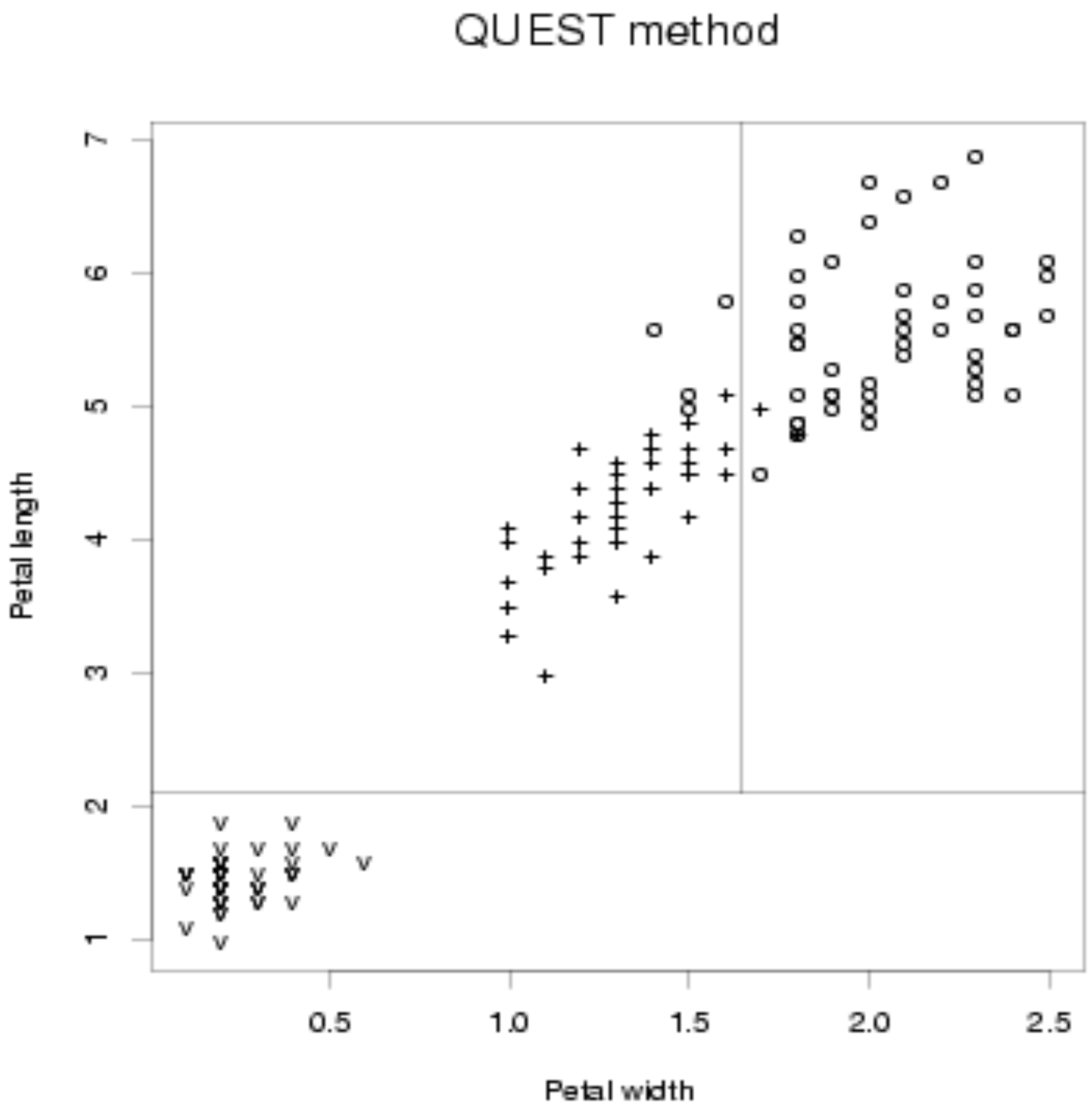
Iris Dataset





Decision Trees: Example




QUEST method
Petal length
Petal width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5

What Are Decision Trees?

- Hierarchical Classifier: breaks complex decision into series of simple decisions
  - each node performs a (single-variable) test to reduce uncertainty
    - For each input variable determine best split
    - Compare among the different variables to select best variable to split on
  - terminal nodes indicate samples (mostly) belonging to the same class (low uncertainty)
- goal: obtain small, shallow tree and low uncertainty (impurity) at the terminals

Decisions for Decision Trees

- How to decide where to split?
  - Goal: have children that are “purer” in class distribution compared to the parent.
  - Mathematical criterion: Entropy, Gini, Ch-squared.
  - Can generalize all three to k-ary splits

- What size tree? (Question of Generalization)
  - Apriori termination criterion
  - grow and prune

- How to handle missing values?
  - Separate group
  - have secondary splitting variable

Decision Tree Packages

- C4.5 (Machine Learning) Uses entropy criterion for split
  - (maximizes information gain)
  - Commercial version is C5.0
- CART (Scientific/Stats): default is gini criterion
- CHAID (marketing/stats): uses Chi-sq; combines variables that are least discriminative,…
- …...

Note: CART also used for regression.

Regression Trees

What is the splitting criterion?

E.g. - Baseball Players Salaries

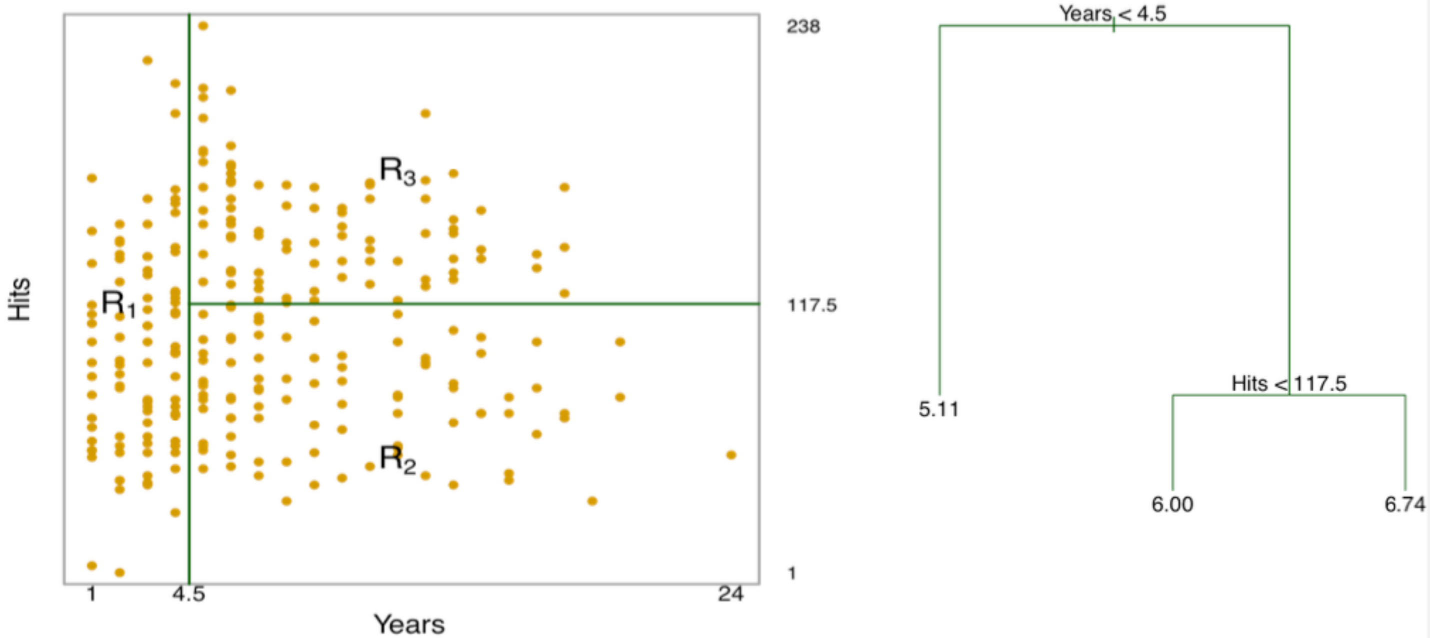


Figure: Regression Tree for Baseball Salaries (in '000, log transformed, so 6 means $402,834

Evaluation of DTs:

- \+ simple, intuitive, often fast, explainable
- \+ integrates feature selection with classification
- \+ can "handle" missing values

- \- often substantially poorer performance
- \- limited: problems with complex decision boundaries, correlated features, continuous variables, …
- \- unstable (partially addressed by bagging/boosting ensemble techniques)

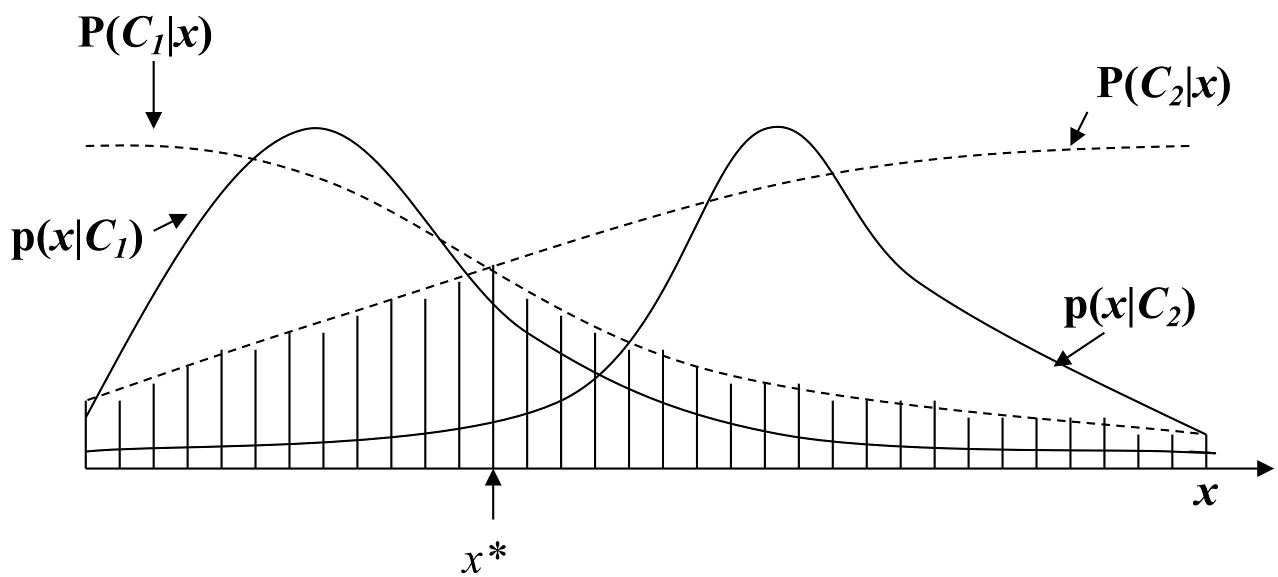
Pattern Recognition Approaches

Founded on Bayes Decision Theory

See Bishop Ch 1.5,

(also KM 5-5.1.2, EA 3-3.4)

Bayes Decision Theory (see Bishop Ch 1.5)



$\mathrm{P}(C_i|x)$ – *a posteriori* probability

$\mathrm{p}(x|C_i)$ – (class conditional) likelihood function

$\mathrm{P}(C_i)$ – class priors

Bayes Classifier

The Bayesian classifier is a parametric method based on

- The *a priori* distributions of classes P(C_i)
- The probability distributions p(x |C_i)
- The *a posteriori* distributions of classes P(C_i|x)

The Bayesian classifier is a MAP classifier (*maximum a posteriori*) : an observed pattern x is classified as class C_i if

$$i = \underset{j=1....K}{\operatorname{argmax}}\{P(C_j \mid x)\}$$

where K is # classes, and

$$P(C_i| x) = \frac{P(C_i)p(x | C_i)}{p(x)}$$

Note: argmax P(C_j|x) = argmax p(C_j , x)

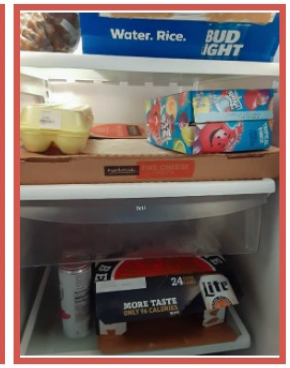
The Bayesian classifier is optimal: it statistically minimizes the error rate

- Catch??

Top Correctly Guessed Trump Refrigerators

86% guessed Trump

85% guessed Trump



85% guessed Trump

84% guessed Trump

Top Correctly Guessed Biden Refrigerators

88% guessed Biden

84% guessed Biden

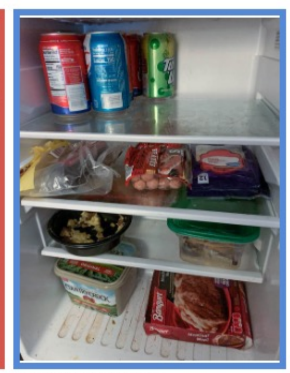
84% guessed Biden

82% guessed Biden

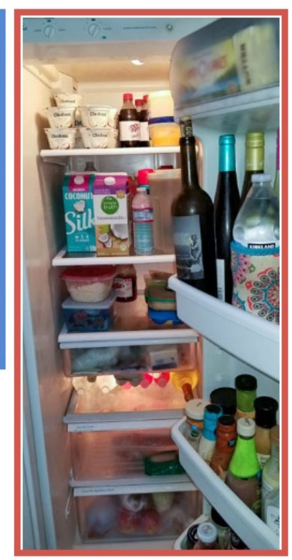
Most Incorrectly Guessed Refrigerators

87% guessed Biden

87% guessed Biden



87% guessed Trump



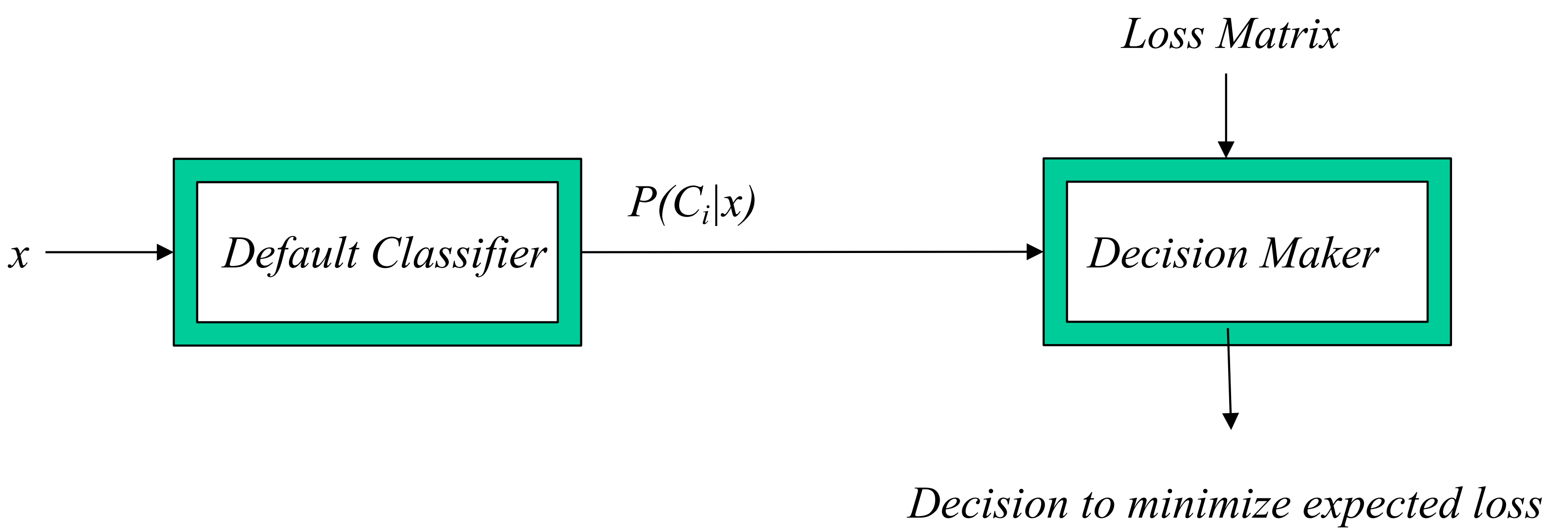
87% guessed Biden

Minimum Expected Loss

- Example: classify medical images as ‘cancer’ or ‘normal’: **costs are asymmetric, hence a LOSS MATRIX of COSTS is needed.**

Decision

$$
\begin{array}{ccc}
 & \text{cancer} & \text{normal} \\
\text{Truth} \quad \begin{array}{l} \text{cancer} \\ \text{normal} \end{array} & \multicolumn{2}{c}{\left(\begin{array}{cc} 0 & 1000 \\ 1 & 0 \end{array}\right)}
\end{array}
$$

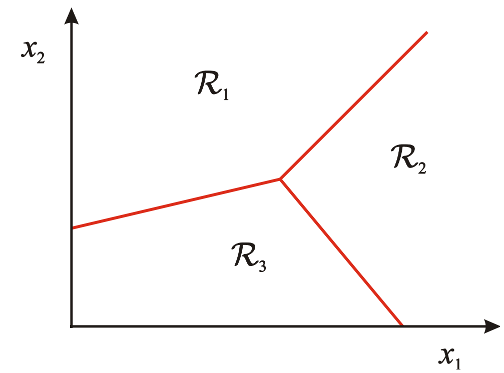


Minimum Expected Loss

- Example: classify medical images as 'cancer' or 'normal': **costs are asymmetric, hence a LOSS MATRIX of COSTS is needed.**

Decision

$$\begin{array}{ccc} & \text{cancer} & \text{normal} \\ \text{Truth} \quad \begin{array}{l}\text{cancer}\\ \text{normal}\end{array} & \multicolumn{2}{c}{\begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix}} \end{array}$$



$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}$$

Hence, given, x, assign class j that minimizes $\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k)$.

("clearly trivial to do once we know the posterior class probabilities", Bishop pg. 42)

<u>Minimum expected cost decision only depends on the Loss Matrix and the Posterior probabilities!</u>
Hence the best label given to a point will only depend on this information: just cycle through the choices for label (j) and pick the one with least cost.

How to Minimize Loss?

- **Example**: consider a **loss matrix**:

		Decision	
		C1	C2
Truth	C1	0	5
	C2	4	-2

Denote P(C1|x) as f(x) for convenience.

For what range of f(x) do we get lower expected loss by assigning to Class 1?

How to Minimize Loss?

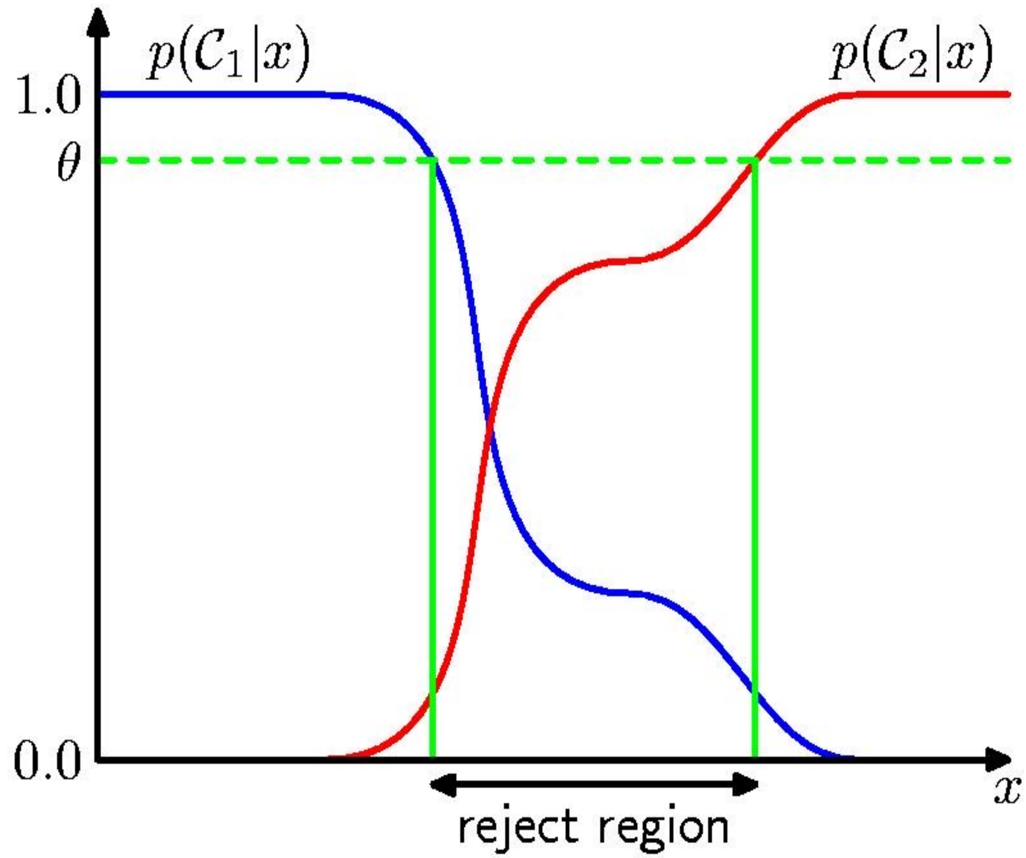
		Decision	
		C1	C2
Truth	C1	0	5
	C2	4	-2

Solution sketch:

Expected loss if x is labeled as class 1 = 0. f(x) + 4 (1-f(x))

Compare with Expected loss if x is labeled as class 2. At the boundary both losses are equal.

Reject Option



Reject Option Example

- **Example**, consider a loss matrix:

		Decision		
		C1	C2	Reject
Truth	C1	0	M_1	R
	C2	M_2	0	R

WLOG (without loss of generality), assume $M_1 > M_2$.

For what range of P(C1|x) am I better off rejecting?

Applying Bayes Decision Theory

- 1. Analytically Solve for Optimal Boundaries via likelihoods.
  - Make assumption about nature of class-conditional distributions
  - Use data to empirically fit distribution for each class.

  Boundary between class i and j means $P(C_i \mid \mathbf{x}) = P(C_j \mid \mathbf{x})$

  e.g. one can show (**See Bishop Sec 4.2.1**) that if each class is normally distributed then boundary is

  a) Quadratic in general (Quadratic Discriminant Analysis or QDA)

  b) Linear (special case when both covariances are the same; get LDA)

  2. Directly try to obtain $P(C_i \mid \mathbf{x})$

Example from KM

- Fig 9.2: QDA left, LDA, right

(a) (b)

Figure 9.1: (a) Some 2d data from 3 different classes. (b) Fitting 2d Gaussians to each class. Generated by code at figures.probml.ai/book1/9.1.

(a) (b)

Figure 9.2: Gaussian discriminant analysis fit to data in Figure 9.1. (a) Unconstrained covariances induce quadratic decision boundaries. (b) Tied covariances induce linear decision boundaries. Generated by code at figures.probml.ai/book1/9.2.

2-D Example from DHS

DHS: Richard O. Duda, Peter E. Hart, and David G. Stork (2001). Pattern Classification. Wiley.)
e-book available from UT Libraries.

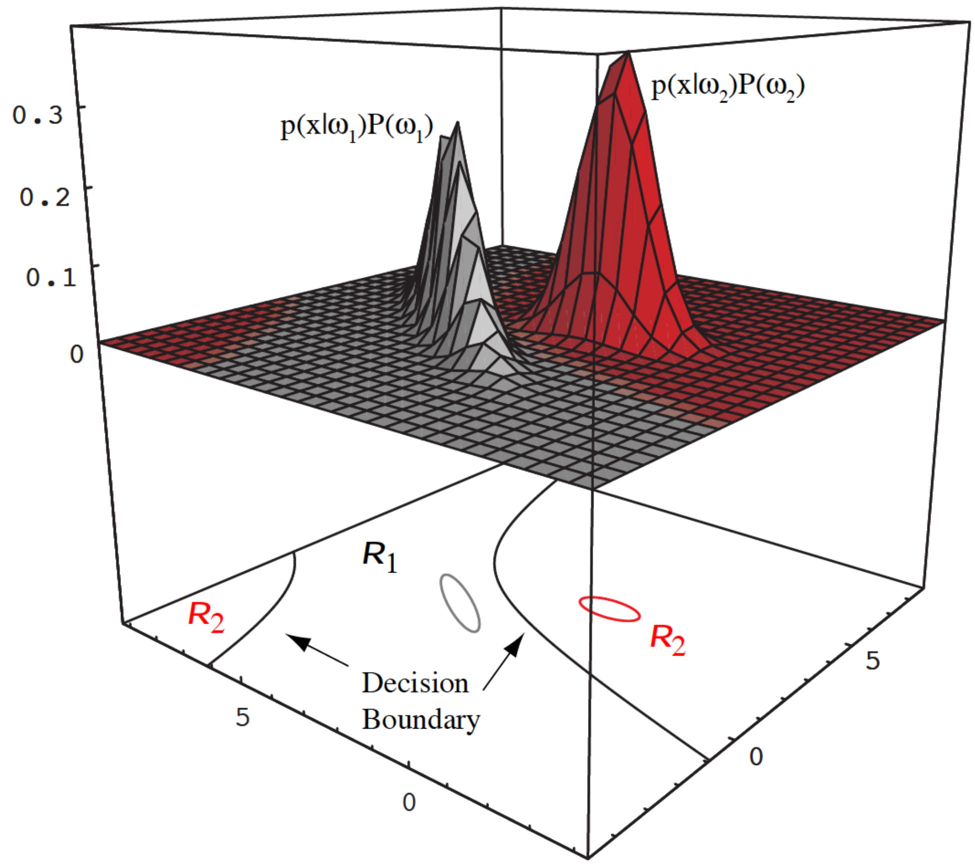


Figure 2.6: In this two-dimensional two-category classifier, the probability densities are Gaussian (with $1/e$ ellipses shown), the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected.

Visuals for Bivariate Gaussians (from DHS)

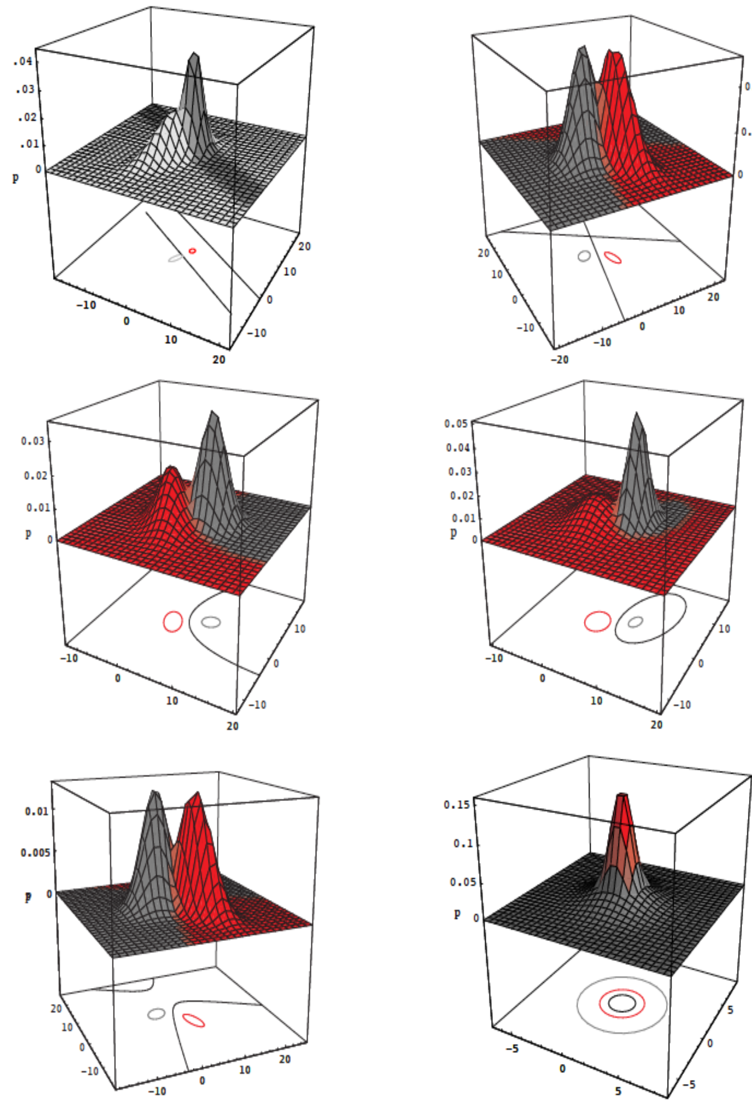


Figure 2.14: Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadratic, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric.

Why Bayes rate is not achieved ?

- Unknown distributions, priors
- Finite training set
  - Leads to estimation errors
- Noisy samples; mislabeled samples
- Missing values
- Symbolic vs. numerical attributes
- (lack of) constraints about problem domains

Bayes rate = function of features used

Measuring Quality of a (Binary) Classifier

Antibody Test Developed for COVID-19 That is Sensitive, Specific and Scalable

SEPTEMBER 11, 2020

Preview (many good videos on both)

- Confusion Matrix
- https://www.youtube.com/watch?v=Kdsp6soqA7o

- ROC Curve
- https://www.dataschool.io/roc-curves-and-auc-explained/

- Text: KM 5.1.3, 5.1.4

Evaluating Results (of any classifier)

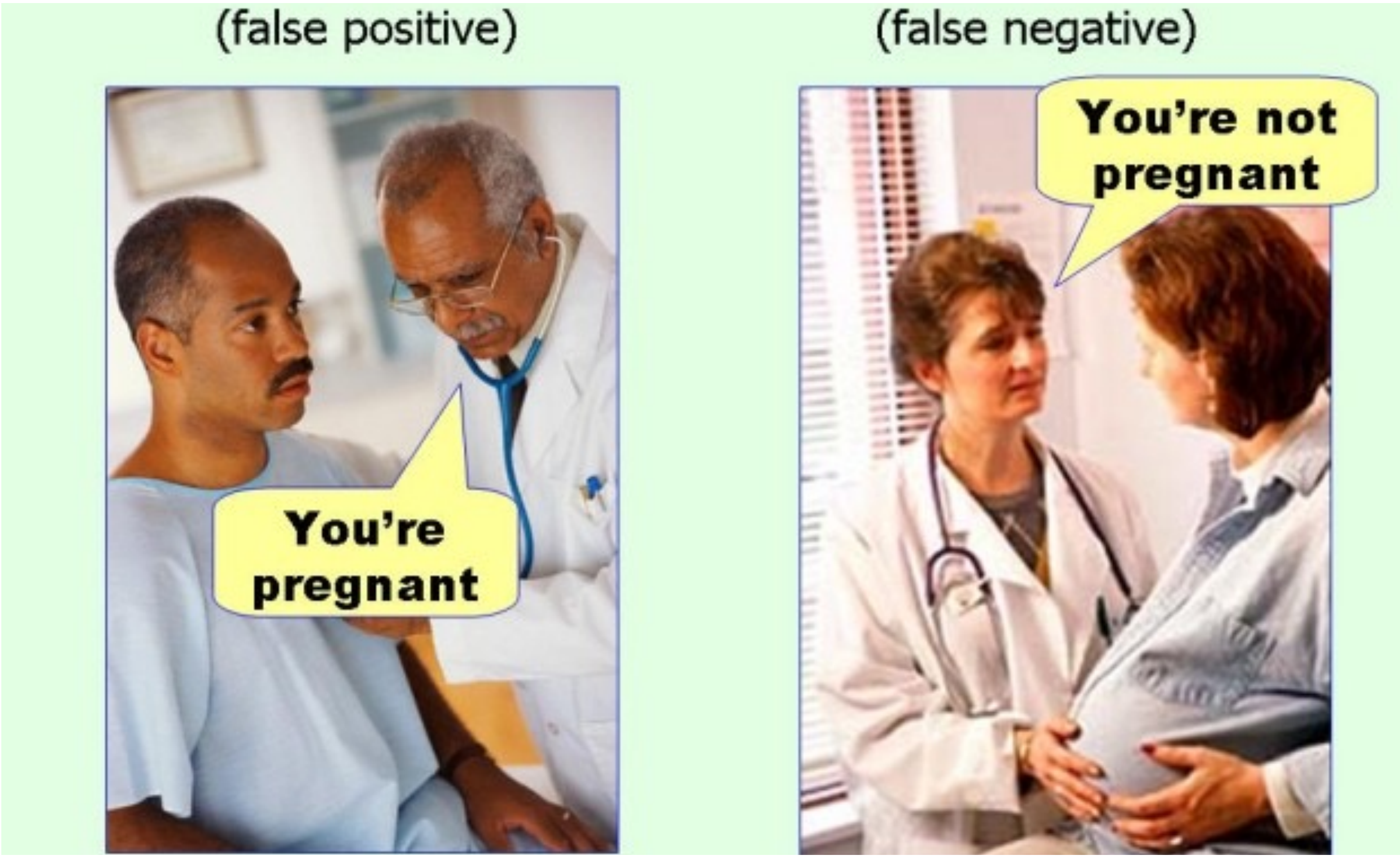
- (correct/mis)-classification rate
  - estimate via holdout, cross-validation,...
- loss or profit
- confusion matrix (video)
  - actual (rows) vs predicted class
  - for two class problems (+ve and –ve class)

  we have:

		Predicted class	
		+ve	-ve
Actual class	+ve	True Positive (TP)	False Negative (FN)
	-ve	False positive (FP)	True Negative (TN)

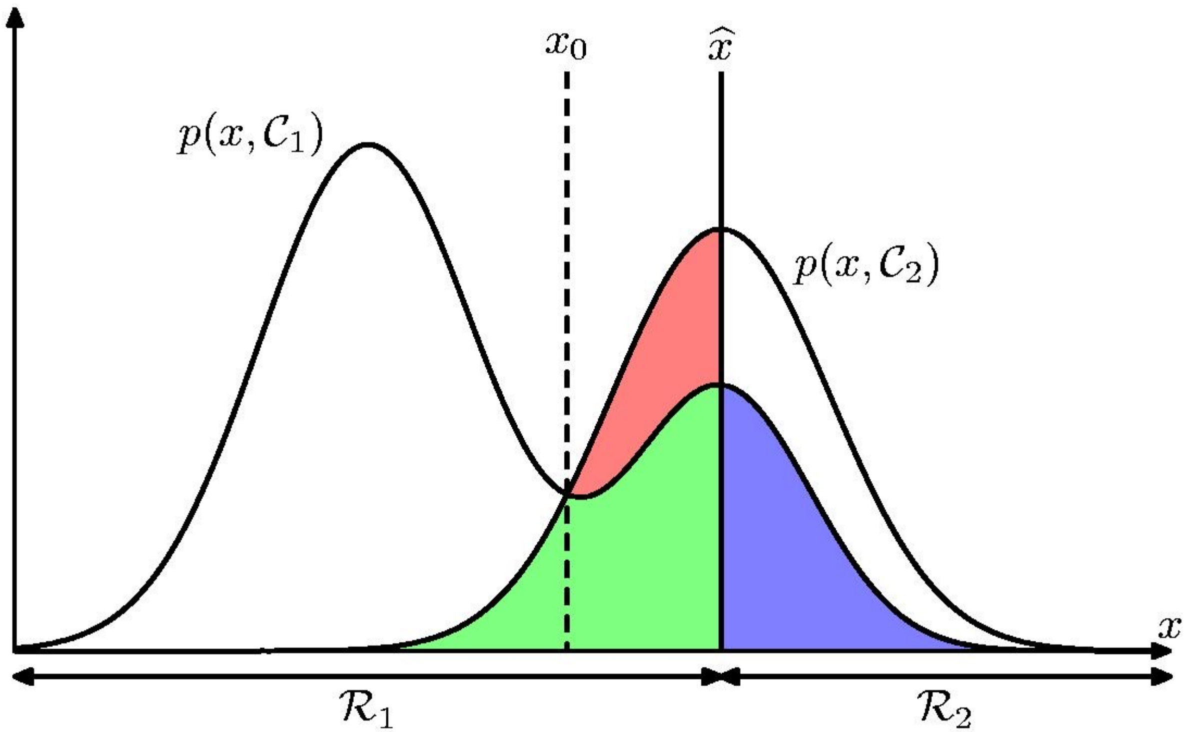
This table has 2 independent parameters

Type I vs Type II Error



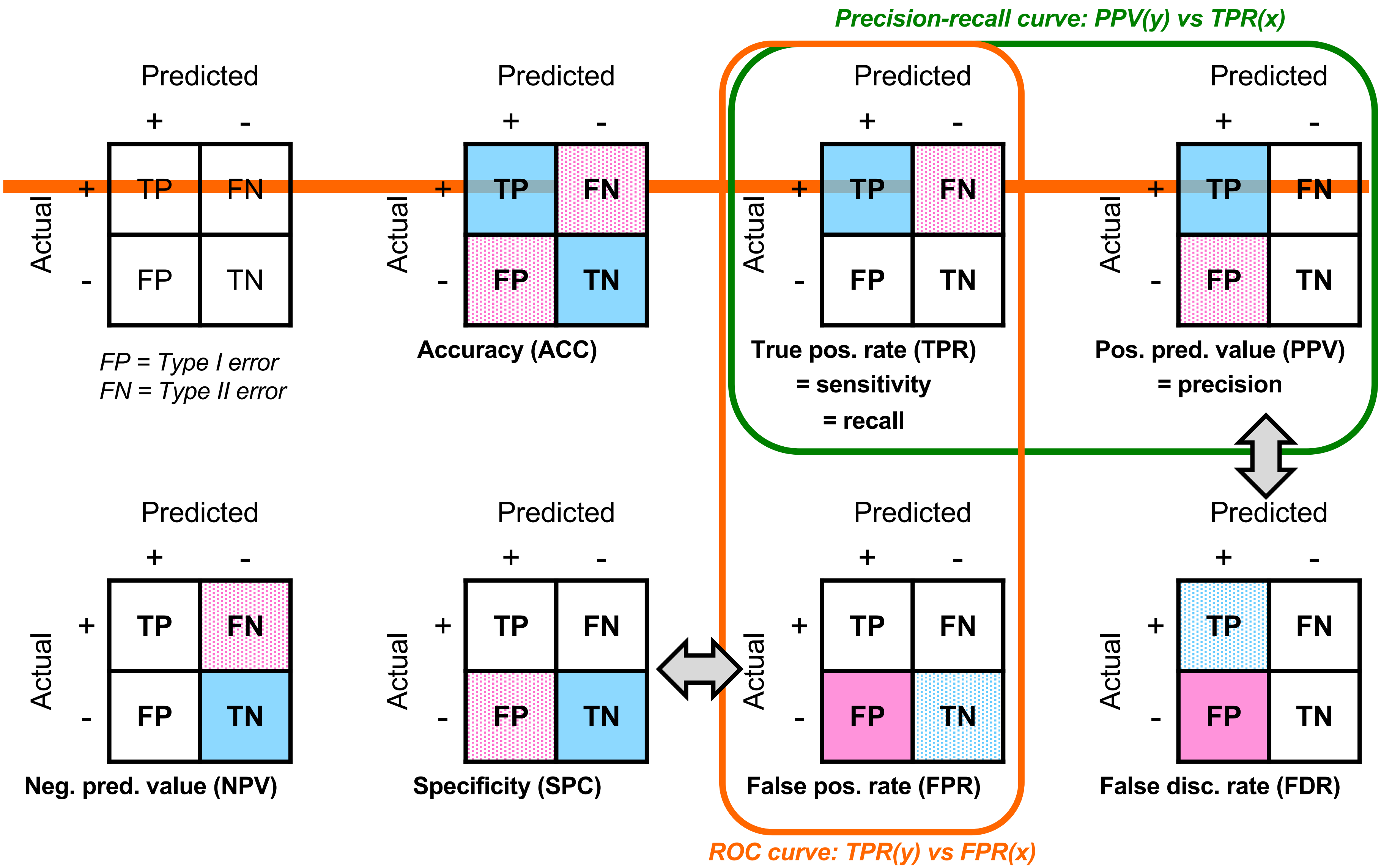
Type III error: when you get the right answer to the wrong question.

Misclassification Rate (Fig 1.24 of Bishop)

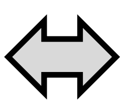


$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}. \end{aligned}$$

R_i denotes the region where data points are assigned to class i



Value: between 0 and 1 (numerator/denominator)
Numerator = solid color shading
Denominator = solid + partial shading


"one minus" relationship

Relationship between ROC and precision-recall:

$$PPV = \frac{P(TPR)}{P(TPR) + N(FPR)} \quad \textit{(ROC to P-R)}$$

$$FPR = \frac{P(1 - PPV)(TPR)}{N(PPV)} \quad \textit{(P-R to ROC)}$$

"P" = # of actual +ves; "N" = # of actual –ves.

Receiver Operating Characteristic (ROC) Charts

- CLASSICAL (Detectors): **% detected (TPR) vs. false alarm rate (FPR)**
  - **(roots in WW II)**

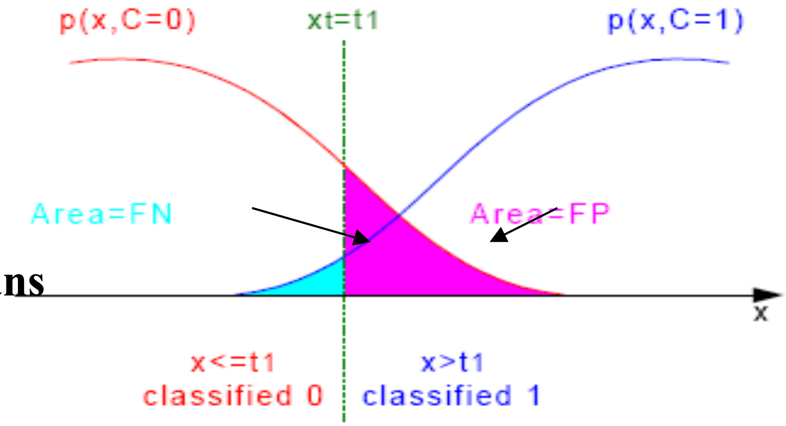
- Medicine: sensitivity vs. (1- specificity) for a range of cutoffs.
  - TP/(TP+FN) vs. FP/(FP+TN)
  - Also talk of Positive Predictive Value (PPV) and Negative Predictive Value (NPV)
- Bayesian Analogues:
  - Sensitivity (Specificity): Accuracy of positive (negative) call conditioned on positive (negative) class.
  - PPV (NPV): Unconditional accuracy of positive (negative) call
- Info. Retrieval: *recall* (fraction of relevant documents retrieved) vs *precision* (fraction of retrieved documents that are relevant)

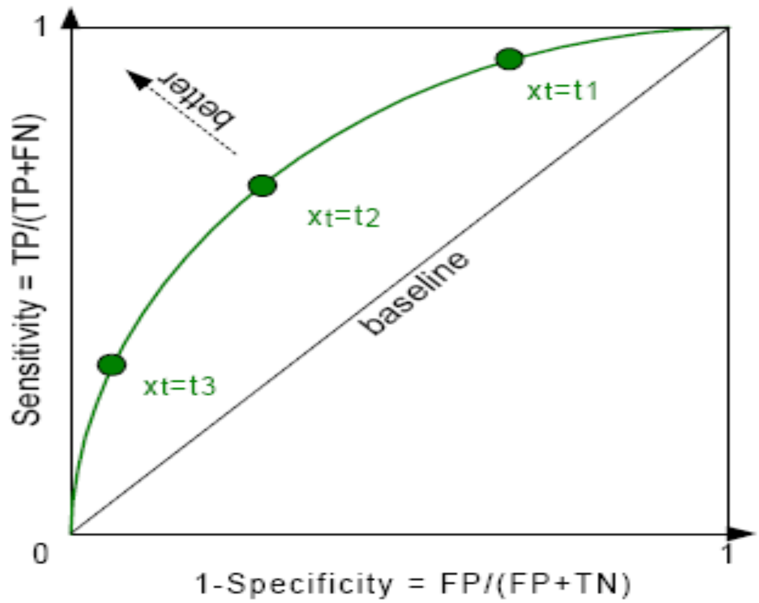
What is a good looking ROC curve? How does it correspond to a Precision-Recall Curve?

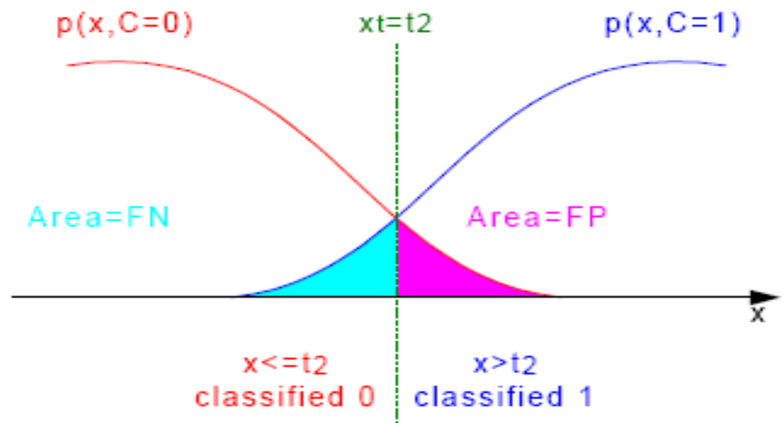
Assume C=1 denotes positive class.

Both entries in a "predicted class" column increase or decrease together.

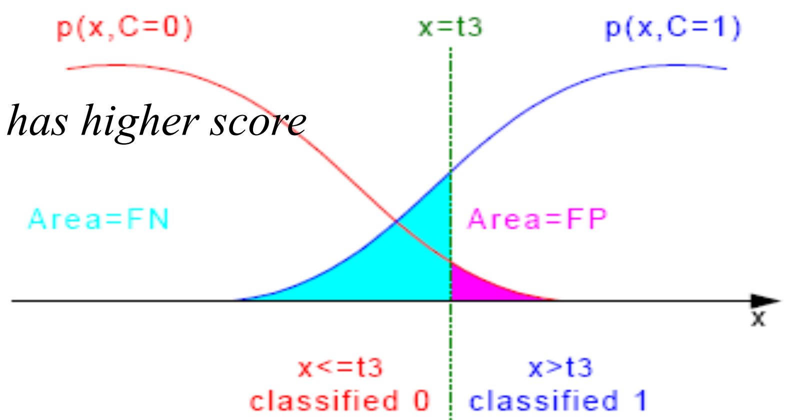
Since total number of samples is same, this means A trade-off between Type I and Type II errors







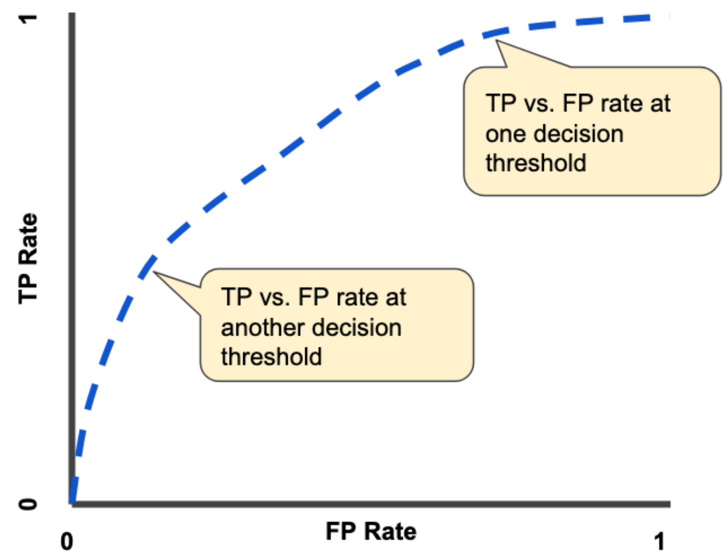
Area under ***ROC curve*** *(AUROC or AUC) = probability that a randomly selected positive example has higher score than a randomly chosen negative example*



See an applet and a tutorial and video

AUC

- [Area under ROC]



- Ideal =1; random =0.5
- One way of interpreting AUC is as the probability that the model

ranks a random positive example more highly than a random negative example. (Details)

Precision Recall Curves

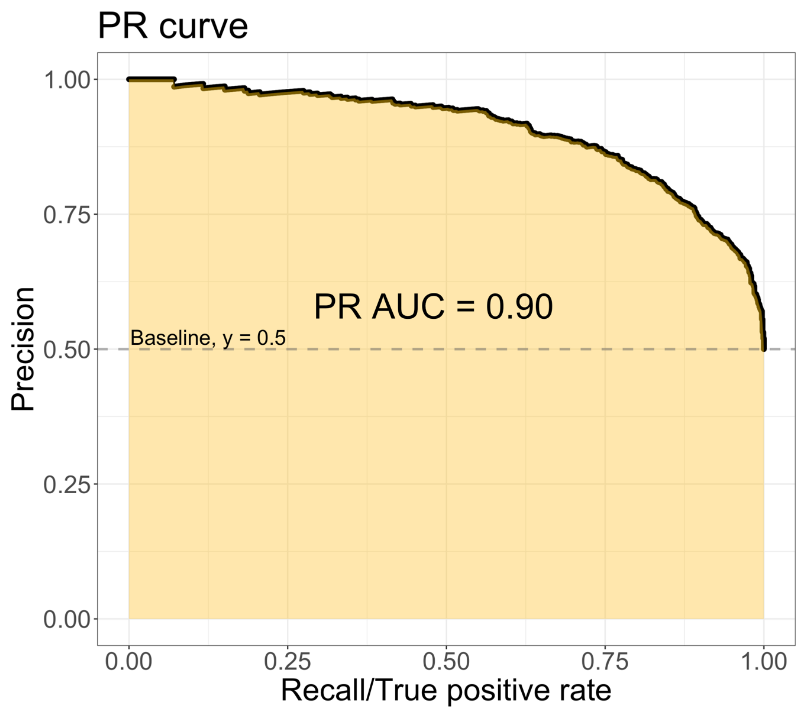


Figure from
https://sinyi-chou.github.io/classification-pr-curve/

Note: Baseline will shift lower if positive class is rarer

(Area under) PR curve is typically a ***better measure for imbalanced classification***.

ROC: treats both classes equally
PR-curve focusses on the positive class.
Also see
https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification

Calibration

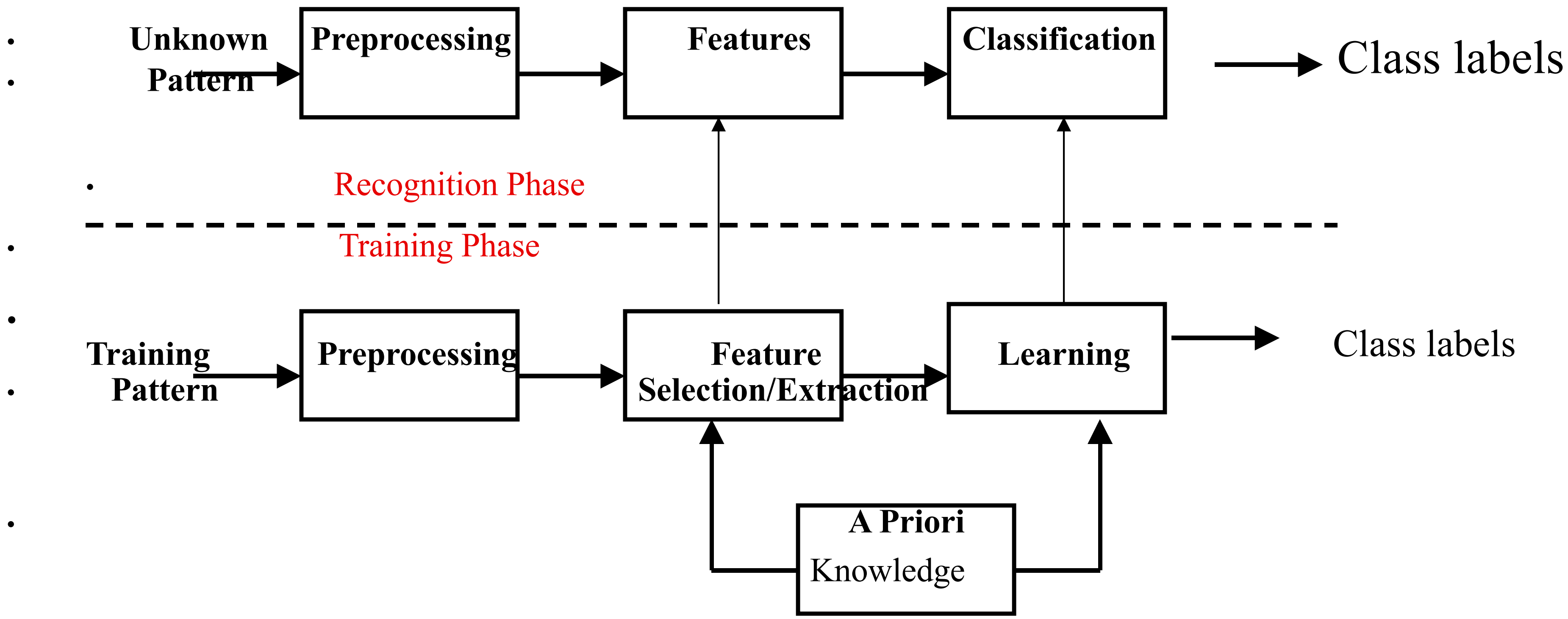
- What is Calibration?
  - Good estimates of posterior probabilities
- When is it important?
- https://scikit-learn.org/stable/auto_examples/calibration/plot_calibration_curve.html#

- What happens to the ROC when all estimates of $P(C|x)$ are multiplied by a constant?

Backups

Training vs. Recognition

- divide given records into training, (validation), and test sets
  - "score" on future data
- true vs. estimated performance



Decision Trees: Evaluation Functions

- Splitting: get "purer" children
  - (class probabilities (p_j' s) estimated empirically)

- Three popular split evaluation functions:
  - entropy : $- \Sigma\, p_j \log p_j$

    tends to prefer attributes with many values
  - gini : $1 - \Sigma\, p_j^{\,2}$

  Both entropy and gini measure impurity at a node
  - use Impurity index: $\Delta\, i(n) = i(n) - p_{left}\, i(n_{left}) - p_{right}\, i(n_{right})$ to evaluate split, where $p_{left,}$ p_{right} are also estimated empirically.

  - Chi-squared contingency table statistic Chi-Sq test with (r-1)x(c-1) degrees of freedom to test if two categorical variables are independent or not
    - Chi-Sq. stats: Σ_{cells} (Observed entry – Expected)^2 / Expected

  - Can generalize all three to k-ary splits

Reasoning About "Actual" Covid Deaths

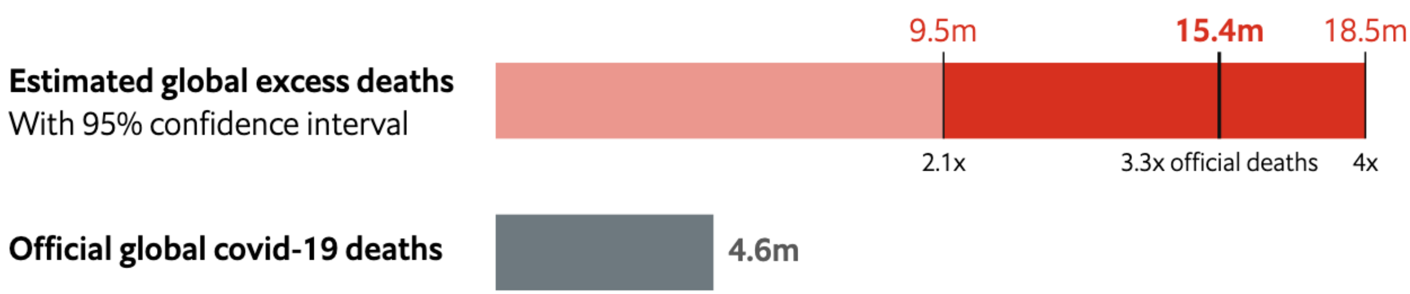
- https://www.economist.com/graphic-detail/coronavirus-excess-deaths-estimates

Covid-19 data

The pandemic's true death toll

Our daily estimate of excess deaths around the world

UPDATED ON SEPTEMBER 12TH



HOW MANY people have died because of the covid-19 pandemic? The answer depends both on the data available, and on how you define "because". Many people who die while infected with SARS-CoV-2 are never tested for it, and do not enter the official totals. Conversely, some people whose deaths have been attributed to covid-19 had other ailments that might have ended their lives on a similar timeframe anyway. And what about people who died of preventable causes during the pandemic, because hospitals full of covid-19 patients could not treat them? If such cases count, they must be offset by deaths that did not occur but would have in normal times, such as those caused by flu or air pollution.

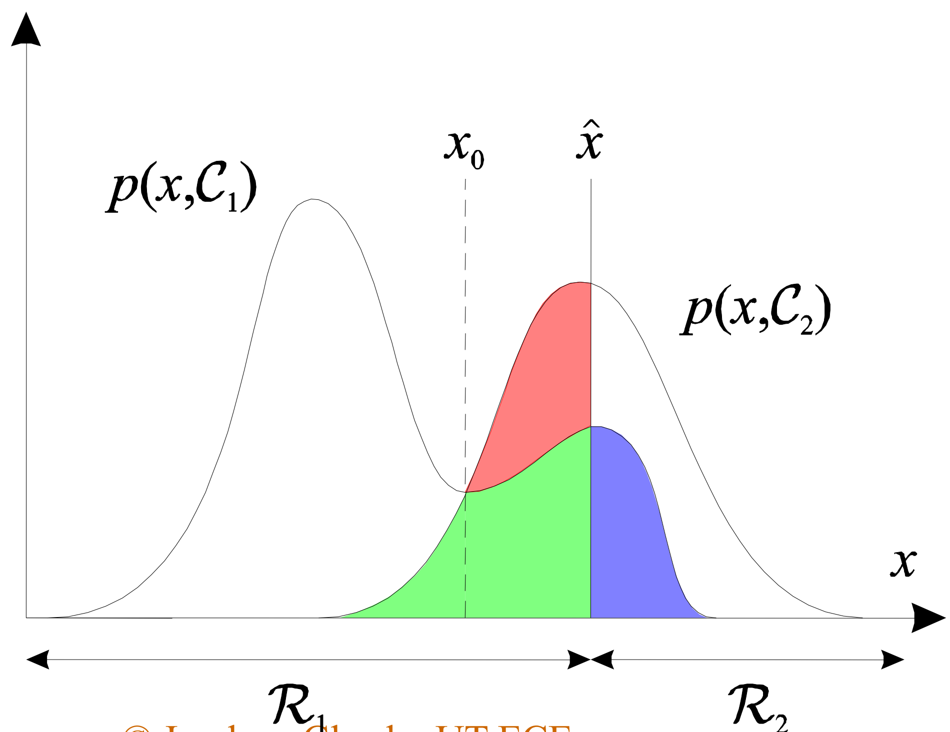
Optimal Decision for Min. Misclassification Rate

- Optimal decision boundary at x_0.
  - P(error) = P(class 2 -> class 1) + P(class 1 -> class 2)

    = (part of green area) + (rest of green area + all of blue area)

So extra error because of non-optimal design = red area.

Ques: show the extra error if boundary is chosen to left of x_0



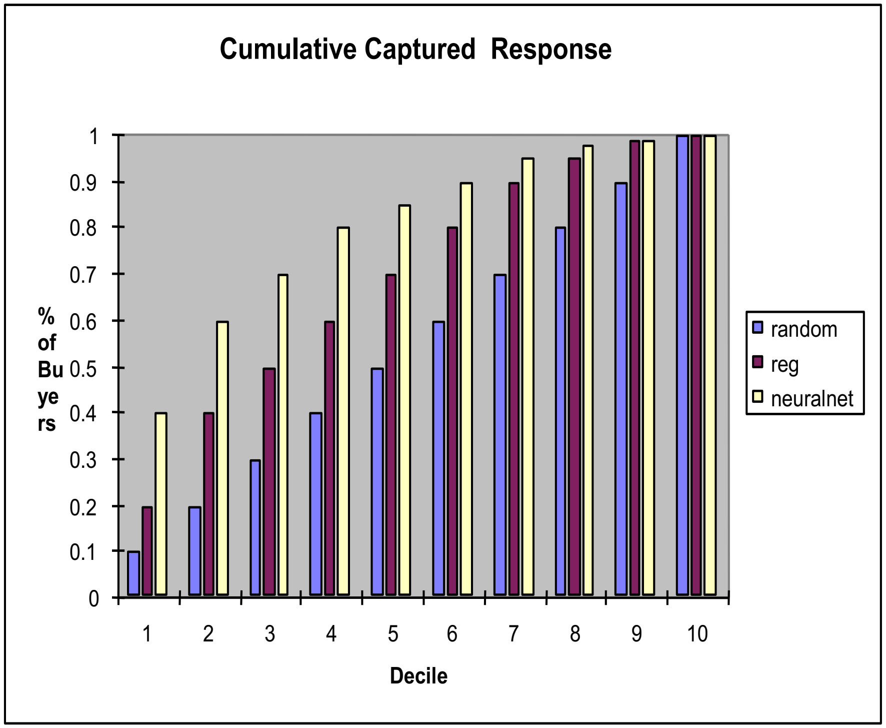
Lift Charts (gains chart)

- sort all observations
  - from highest expected profit to lowest expected profit (non-binary).
  - by the posterior probabilities of the target event (binary targets)
- group into deciles and plot
  - typically, cumulative plot of "% of target events in selection"
    - often normalized by "% in a random selection"
      e.g.: top decile provides a lift of 2.5

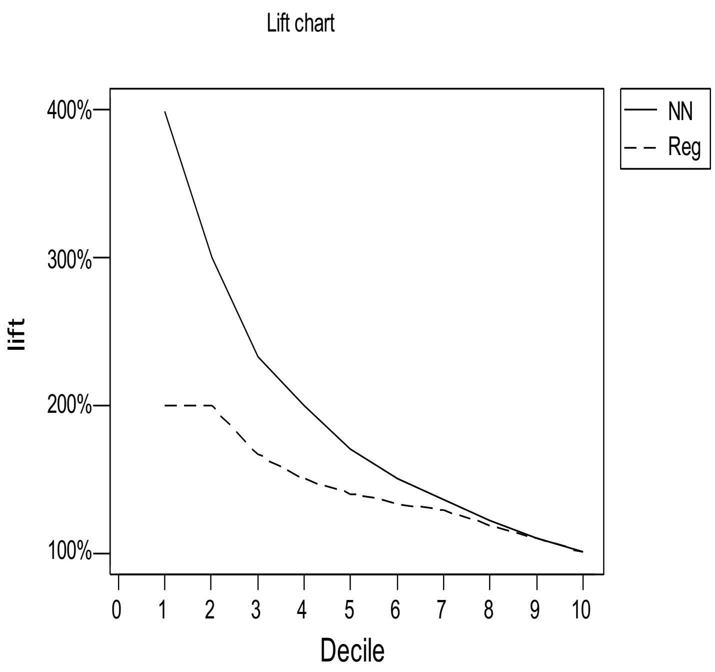
      Often only top 10% or top 25% is of interest

Lift Example (direct marketing)

Cumulated Lift Chart



Lift Chart



Example: Logistic Regression can capture 50% of the buyers by mailing to 30% of target audience; neural net can capture 70% of the buyers by mailing to 30% of the audience.

Cumulative Lift vs. ROC

- Hint: what are expressions for the X and Y axes for each graph?

Effect of Resampling (for handling class imbalance)

- Original binary classification problem has $P(C1) < 0.5$
  - Bayes decision boundary at $P(C1|x) = 0.5$

- Resample C1 and C2, to get equal priors. Call these two datasets C3 and C4 respectively
  - (only priors have changed, class conditional likelihoods have not).
- Show that the original decision boundary will be obtained at $P(C3|x) = 1 - P(C1)$
  - Need a higher threshold.

Why Separate Inference and Decision?

- **<u>Inference:</u> estimate the P(Ci|x) terms**
- **<u>Decision:</u> allocate a given x to a specific class.**

- **Minimizing risk (loss matrix may change over time)**
- **Reject option**
- **Changed class priors in scoring data**
- **Combining models (later in the course)**

Generative vs. Discriminative Approaches

- Generative: separately model class-conditional densities and priors
  (e.g. LDA)

- Discriminative: try to obtain class boundaries directly
  Through heuristic or through directly estimating posterior probabilities
  Just predict class label.
  (e.g. Decision Trees)

  Pros and cons?