HST.582J / 6.555J / 16.456J Biomedical Signal and Image Processing
Spring 2007

Automated Decision Making Systems

Probability, Classification, Model Estimation

Information and Statistics

One the use of statistics:

"There are three kind of lies: lies, damned lies, and statistics"
- Benjamin Disraeli (popularized by Mark Twain)

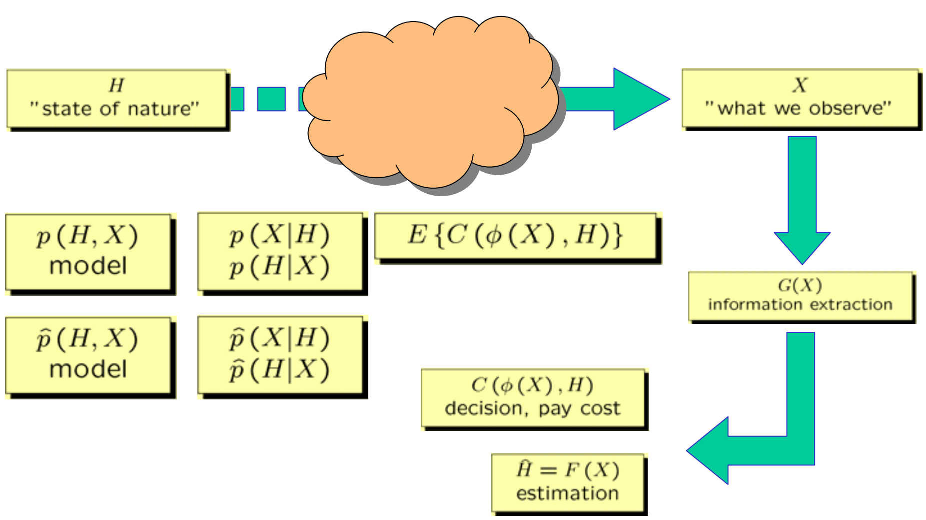
On the value of information:

"And when we were finished renovating our house, we had only $24.00 left in the bank only because the plumber didn't know about it."
- Mark Twain (from a speech paraphrasing one of his books)

Elements of Decision Making Systems

1. Probability
   - A *quantitative* way of modeling uncertainty.
2. Statistical Classification
   - application of probability models to inference.
   - incorporates a notion of optimality
3. Model Estimation
   - we rarely (OK never) know the model beforehand.
   - can we estimate the model from labeled observations.

Problem Setup

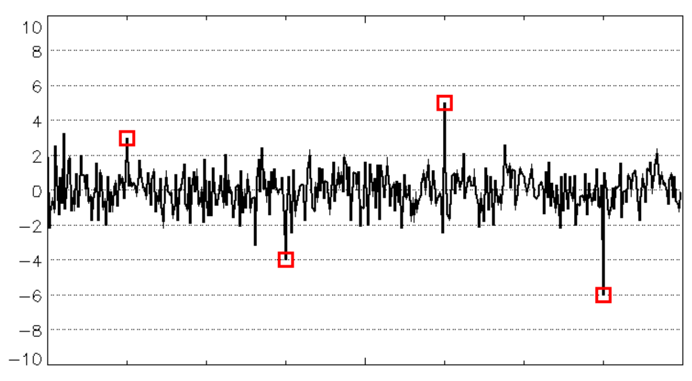


Concepts

- In many experiments there is some element of randomness the we are unable to explain.
- Probability and statistics are mathematical tools for reasoning in the face of such uncertainty.
- They allow us to answer questions *quantitatively* such as
  - Is the signal present or not?
    - Binary : YES or NO
  - How certain am I?
    - Continuous : Degree of confidence
- We can design systems for which
  - Single use performance has an element of uncertainty
  - Average case performance is predictable

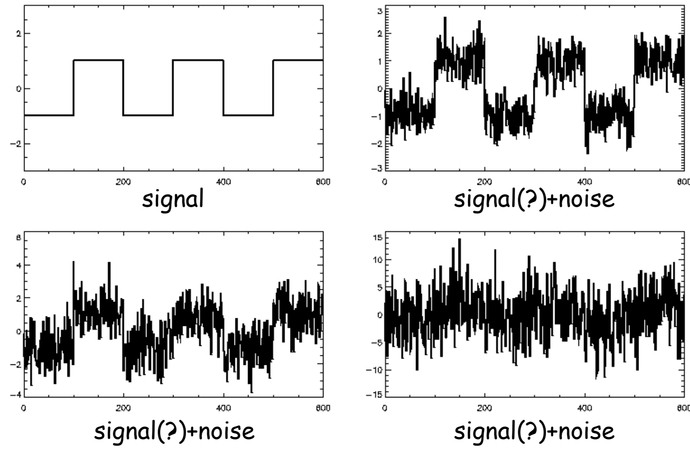
Anomalous behavior (example)

- How do quantify our belief that these are anomalies?
- How might we detect them automatically?



Detection of signals in noise

- In which of these plots is the signal present?
- Why are we more certain in some cases than others?



Coin Flipping

- Fairly simple probability modeling problem
  - Binary hypothesis testing
  - Many decision systems come down to making a decision on the basis of a biased coin flip (or N-sided die)

Bayes' Rule

- Bayes' rule plays an important role in classification, inference, and estimation.

$$
\begin{aligned}
P(AB) &= P(A|B)\,P(B) \\
&= P(B|A)\,P(A)
\end{aligned}
$$

$$
\begin{aligned}
P(A|B) &= \frac{P(B|A)\,P(A)}{P(B)} \\
&= \frac{P(BA)}{P(B)}
\end{aligned}
$$

$$
\begin{aligned}
P(B|A) &= \frac{P(A|B)\,P(B)}{P(A)} \\
&= \frac{P(AB)}{P(A)}
\end{aligned}
$$

- A useful thing to remember is that **conditional** probability relationships can be derived from a Venn diagram. **Bayes' rule** then arises from straightforward algebraic manipulation.

Heads/Tails Conditioning Example

- If I flip two coins and tell you **at least** one of them is "heads" what is the probability that **at least** one of them is "tails"?
- The events of interest are the set of outcomes where ***at least*** one of the results is a head.
- The point of this example is two-fold
  - Keep track of your sample space and events of interest.
  - Bayes' rule tells how to incorporate information in order to adjust probability.

		2nd flip	
		H	T
1st flip	H	HH	HT
	T	TH	TT

Heads/Tails Conditioning Example

- The probability that ***at least*** one of the results is heads is $\frac{3}{4}$ by simple counting.
- The probability that both of the coins are heads is $\frac{1}{4}$

$$\begin{aligned} A &= \text{the "other" coin is heads} \\ B &= \text{at least one of the coins is heads} \\ AB &= \text{both of the coins are heads} \\ P(A|B) &= \frac{P(BA)}{P(B)} \end{aligned}$$

- The chance of winning is 1 in 3
- Equivalently, the odds of winning are 1 to 2

		2nd flip	
		H	T
1st flip	H	HH	HT
	T	TH	TT

		2nd flip	
		H	T
1st flip	H	HH	HT
	T	TH	TT

Defining Probability (Frequentist vs. Axiomatic)

The *probability* of an event is the number of times we expect a specific outcome relative to the number of times we conduct the experiment.

Define:

- N: the number of trials
- N_A, N_B : the number of times events **A** and **B** are observed.
- Events **A** and **B** are mutually exclusive (i.e. observing one precludes observing the other).

Empirical definition:

- Probability is defined as a limit over observations

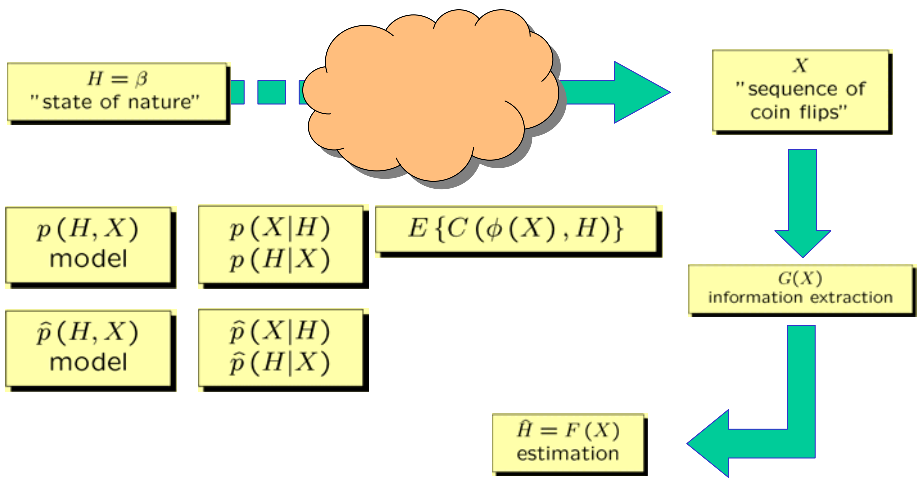
$$\begin{aligned} P\{A\} &= \lim_{N\to\infty}\left(\frac{N_A}{N}\right) \\ P\{B\} &= \lim_{N\to\infty}\left(\frac{N_B}{N}\right) \\ P\{A+B\} &= \lim_{N\to\infty}\left(\frac{N_A+N_B}{N}\right) \end{aligned}$$

Axiomatic definition:

- Probability is derived from its properties

$$\begin{aligned} &0 \le P\{A\}, P\{B\} \le 1 \\ &P\{\text{the certain event}\} = 1 \\ &P\{A+B\} = P\{A\} + P\{B\} \end{aligned}$$

Estimating the Bias of a Coin (Bernoulli Process)



4 out of 5 Dentists…

- What does this statement mean?
- How can we attach meaning/significance to the claim?

- An example of a frequentist vs. Bayesian viewpoint
  - The difference (in this case) lies in:
    - The assumption regarding how the data is generated
    - The way in which we can express certainty about our answer
  - Asympotitically (as we get more observations) they both converge to the same answer (but at different rates).

Sample without Replacement, Order Matters

Begin with N empty boxes

- each term represents the number of different choices we have at each stage

$$N \times (N-1) \times (N-2) \times \cdots \times (N-k+1)$$

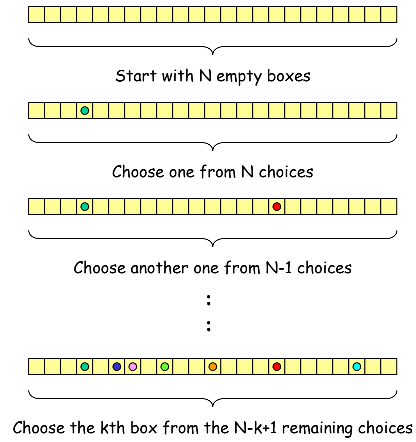
- this can be re-written as

$$\frac{N \times (N-1) \times (N-2) \times \cdots \times 2 \times 1}{(N-k) \times (N-k-1) \times (N-2) \times \cdots \times 2 \times 1}$$

- and then “simplified” to

$$\frac{N!}{(N-k)!}$$

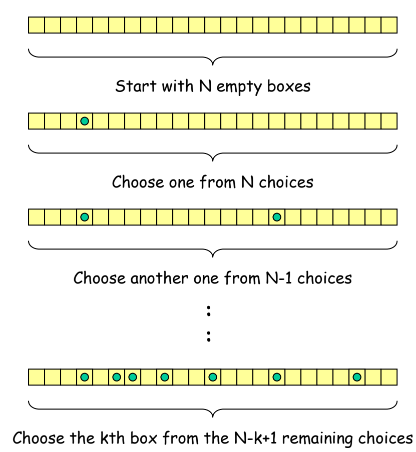
At left: color indicates the *order* in which we filled the boxes. Any sample which fills the same boxes, but has a different color in any box (there will be at least 2) is considered a *different* sample.



Sample without Replacement, Order doesn’t Matter

- The sampling procedure is the same as the previous *except* that we don’t keep track of the colors.
- The number of sample draws with the same filled boxes is equal to the number of ways we can re-order (permute) the colors.
- The result is to reduce the total number of draws by that factor.

$$\frac{N!}{(N-k)!} \rightarrow \frac{N!}{(N-k)!k!} = \binom{N}{k}$$



Cumulative Distributions Functions (PDFs)

- **cumulative distribution function (CDF)** divides a continuous *sample space* into two *events*

$$P_X(x) = \Pr\{X \leq x\} \qquad 1 - P_X(x) = \Pr\{X > x\}$$

- It has the following properties

$$\begin{aligned} P_X(-\infty) &= 0 \\ P_X(\infty) &= 1 \\ 0 &\leq P_X(x) \leq 1 \\ P_X(x+\Delta) &\geq P_X(x) \quad ; \quad \Delta \geq 0 \end{aligned}$$

Probability Density Functions (PDFs)

- **probability density function (PDF)** is defined in terms of the **CDF**

$$\begin{aligned} P_X(x) &= \int_{-\infty}^{x} p_x(u)\,du \\ p_X(x) &= \frac{\partial}{\partial x} P_X(x) \end{aligned}$$

- Some properties which follow are:

$$\begin{aligned} \int_{-\infty}^{\infty} p_x(u)\,du &= 1 \\ p_X(x) &\geq 0 \end{aligned}$$

Expectation

- Given a function of a random variable (i.e. g(X)) we define it's **expected** value as:

$$
\begin{aligned}
E\{g(X)\} &= \sum_{i=1}^{N} g(x_i) p_x(x_i) \\
&= \int_{\Omega_X} g(u) p_x(u)\, du
\end{aligned}
$$

- For the mean, variance, and entropy (continous examples):

$g(X)$	statistic
X	mean
$(X - E\{X\})^2$	variance
$-\log(p_x(X))$	entropy

- Expectation is linear (see variance example once we've defined ***joint density function*** and ***statistical independence***)

$$
\begin{aligned}
E\{\alpha f(x) + \beta g(x)\} &= \alpha E\{f(x)\} \\
&\quad + \beta E\{g(x)\}
\end{aligned}
$$

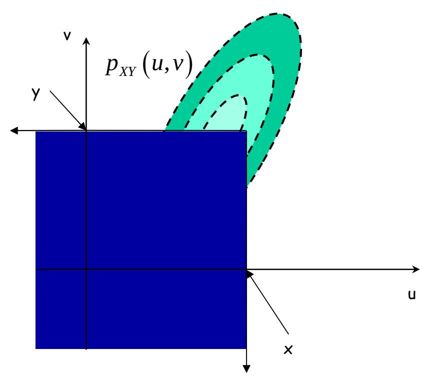
- Expectation is with regard to *ALL* random variables within the arguments.
  - This is important for multi-dimensional and joint random variables.

Multiple Random Variables (Joint Densities)

We can define a density over multiple random variables in a similar fashion as we did for a single random variable.

1. We define the probability of the event $\{X \leq x \text{ AND } Y \leq y\}$ as a function of x and y.
2. The density is the function we integrate to compute the probability.

$$
\begin{aligned}
P_{XY}(x,y) &= \Pr\{X \leq x \text{ AND } Y \leq y\} \\
&= \int_{-\infty}^{x} \int_{-\infty}^{y} p_{xy}(u,v)\, du dv \\
p_{XY}(x,y) &= \frac{\partial^2}{\partial x \partial y} P_{XY}(x,y)
\end{aligned}
$$



$P_{XY}(x,y)$ is the area under the curve integrated over shaded region for a given $\{x,y\}$

Conditional Density

Given a joint density or mass function over two random variables we can define the conditional density similar to conditional probability from Venn diagrams

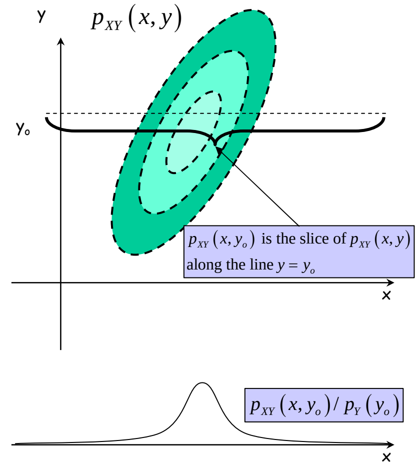
$$p_{x|y}(x|y) = \frac{p_{xy}(x,y)}{p_y(y)}$$

This is, it is not of practical use unless we condition on Y equal to a ***value*** versus letting it remain a variable (creating an actual density)

$$p_{x|y}(x|y=y_o) = \frac{p_{xy}(x,y_o)}{p_y(y_o)}$$

We also get the following relationship

$$\begin{aligned} p_{x|y}(x|y)\,p_y(y) &= p_{xy}(x,y) \\ &= p_{y|x}(y|x)\,p_x(x) \end{aligned}$$



$p_{x|y}(x|Y=y_o)$ is $p_{xy}(x,y_o)$ (a slice of the joint density, normalized by $p_y(y_o)$ so that it integrates to unity (i.e. it is a valid density)

Bayes' Rule

- For continuous random variables, Bayes' rule is essentially the same (again just an algebraic manipulation of the definition of a conditional density).

$$p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)\,p_X(x)}{p_Y(y)}$$

- This relationship will be very useful when we start looking at classification and detection.

Binary Hypothesis Testing (Neyman-Pearson)
(and a "simplification" of the notation)

- 2-Class problems are equivalent to the binary hypothesis testing problem.

$$\begin{aligned} H_1 &: x \sim p_{X|H_1}(x|H_1 \text{ is true}) \\ H_0 &: x \sim p_{X|H_0}(x|H_0 \text{ is true}) \end{aligned}$$

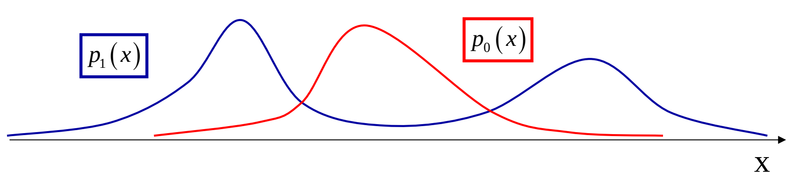
The goal is *estimate* which Hypothesis is true (i.e. from which class our sample came from).

- A minor change in notation will make the following discussion a little simpler.

$$\begin{aligned} p_1(x) &= p_{X|H_1}(x|H_1 \text{ is true}) \\ p_0(x) &= p_{X|H_0}(x|H_0 \text{ is true}) \end{aligned}$$

Probability density models for the measurement x depending on which hypothesis is in effect.

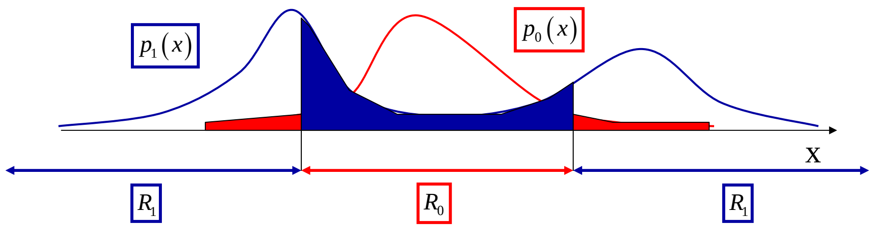
Decision Rules



- Decision rules are functions which map measurements to choices.
- In the binary case we can write it as

$$\phi(x) = \begin{cases} 1 & ; \; x \in R_1 \\ 0 & ; \; x \in R_0 \end{cases}$$

Error Types



- There are 2 types of errors
- A "miss"

E_M: X falling in R_0 **AND** H_1 being correct

- A "false alarm"

E_F: X falling in R_1 **AND** H_0 being correct

Binary Hypothesis Testing (Bayesian)

- 2-Class problems are equivalent to the binary hypothesis testing problem.

$$\begin{aligned} H_1 &: x \sim p_{X|H_1}(x|H_1 \text{ is true}) \\ H_0 &: x \sim p_{X|H_0}(x|H_0 \text{ is true}) \end{aligned}$$

The goal is *estimate* which Hypothesis is true (i.e. from which class our sample came from).

- A minor change in notation will make the following discussion a little simpler.

$$\begin{aligned} P_1 &= \Pr(H = H_1) \\ P_0 &= \Pr(H = H_0) \end{aligned}$$

} Prior probabilities of each class

$$\begin{aligned} p_1(x) &= p_{X|H_1}(x|H_1 \text{ is true}) \\ p_0(x) &= p_{X|H_0}(x|H_0 \text{ is true}) \end{aligned}$$

} Class-conditional probability density models for the measurement x

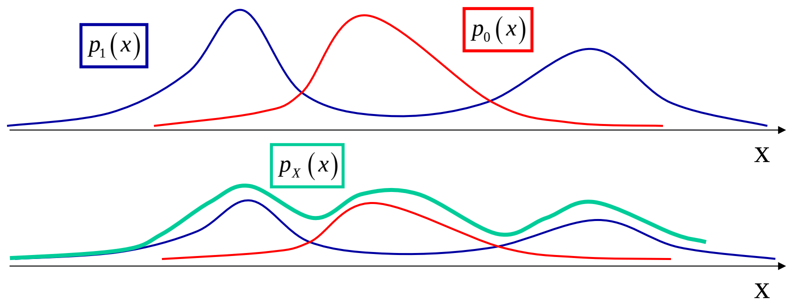
Marginal density of X

$$p_x(x) = P_1 p_1(x) + P_0 p_0(x)$$

Conditional probability of the hypothesis H_i given X

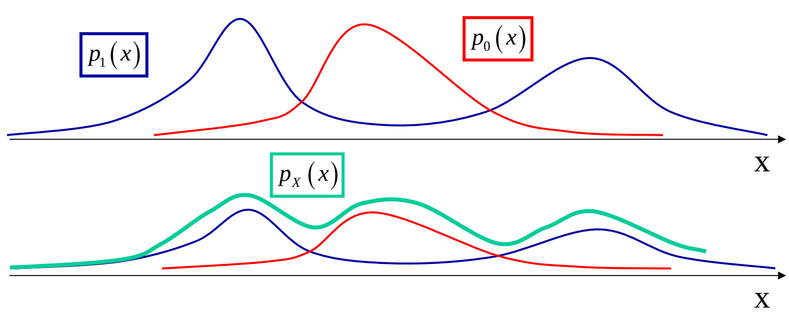
$$\begin{aligned} P_{H_i|x}(H_i|x) &= \frac{P_i p_i(x)}{p_x(x)} \\ &= \frac{P_i p_i(x)}{P_1 p_1(x) + P_0 p_0(x)} \end{aligned}$$

A Notional 1-Dimensional Classification Example



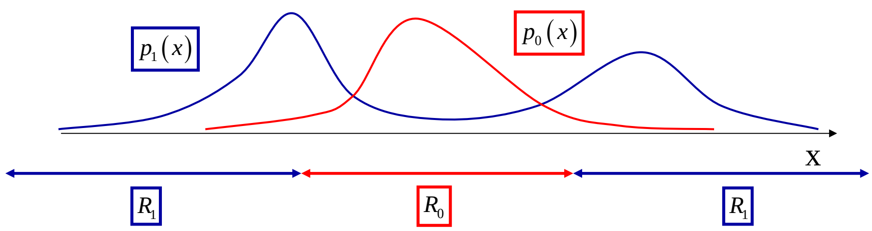
- So given observations of *x*, how should select our best guess of H_i?
- Specifically, what is a good criterion for making that assignment?
- Which H_i should we select before we observe *x*.

Bayes Classifier



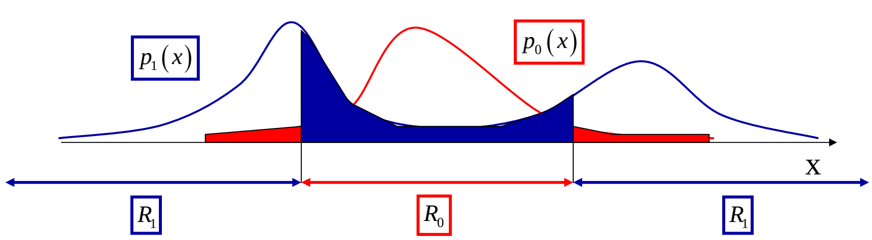
- A reasonable criterion for guessing values of H given observations of X is to minimize the probability of error.
- The classifier which achieves this minimization is the Bayes classifier.

Probability of Misclassification



- Before we derive the Bayes' classifier, consider the probability of misclassification for an **arbitrary** classifier (i.e. decision rule).
  - The first step is to assign regions of X, to each class.
  - An error occurs if a sample of x falls in R_i and we assume hypothesis H_j.

Probability of Misclassification



- An error is comprised of two events

E_1: X falling in R_0 **AND** H_1 being correct

E_0: X falling in R_1 **AND** H_0 being correct

- These are *mutually exclusive* events so their joint probability is the sum of their individual probabilities

$$\begin{aligned} P_E &= \Pr\{E_1\} + \Pr\{E_0\} \\ &= P_1 \Pr\{X \in R_0 | H_1\} + P_0 \Pr\{X \in R_1 | H_0\} \\ &= P_1 \int_{R_0} p_1(x)\, dx + P_0 \int_{R_1} p_0(x)\, dx \end{aligned}$$

Minimum Probability of Misclassification

- So now let's choose regions to minimize the probability of error.

$$\begin{aligned} P_E &= P_1 \int_{R_0} p_1(x)\,dx + P_0 \int_{R_1} p_0(x)\,dx \\ &= P_1 \left(1 - \int_{R_1} p_1(x)\,dx\right) + P_0 \int_{R_1} p_0(x)\,dx \\ &= P_1 + \int_{R_1} \left(\underbrace{P_0 p_0(x)}_{\geq 0} - \underbrace{P_1 p_1(x)}_{\geq 0} \right) dx \end{aligned}$$

- In the second step we just change the region over which integrate for one of the terms (these are complementary events).
- In the third step we collect terms and note that all underbraced terms in the integrand are non-negative.
- If we want to choose regions (remember choosing region 1 effectively chooses region 2) to minimize P_E then we should set region 1 to be such that the integrand is negative.

Minimum Probability of Misclassification

- Consequently, for minimum probability of misclassification (which is the Bayes error), R_1 is defined as

$$R_1 = \{x : P_1 p_1(x) > P_2 p_2(x)\}$$

- R_2 is the complement. The boundary is where we have equality.
- Equivalently we can write the condition as when the likelihood ratio for H_1 vs H_0 exceeds the PRIOR odds of H_0 vs H_1

$$R_1 = \left\{ x : \frac{p_1(x)}{p_0(x)} > \frac{P_0}{P_1} \right\}$$

Risk Adjusted Classifiers

Suppose that making one type of error is more of a concern than making another. For example, it is worse to declare H_1 when H_2 is true then vice versa.

- This is captured by the notion of "cost".

$$C_{ij} = \text{cost of declaring } H_i \text{ when } H_j \text{ is correct}$$

- In the binary case this leads to a cost matrix.

$$\text{declared hypothesis} \left\{ \underbrace{\begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix}}_{\text{correct hypothesis}} \right.$$

- The Risk Adjusted Classifier tries to minimize the expected "cost"

Derivation

- We'll simplify by assuming that $C_{11}=C_{22}=0$ (there is zero cost to being correct) and that all other costs are positive.
- Think of cost as a piecewise constant function of X.
- If we divide X into decision regions we can compute the expected cost as the cost of being wrong times the probability of a sample falling into that region.

$$\begin{aligned} E\{C(x,H)\} &= \int_{R_0} C_{01}P_1 p_1(x)\,dx + \int_{R_1} C_{10}P_0 p_0(x)\,dx \\ &= C_{01}P_1\left(1 - \int_{R_1} p_1(x)\,dx\right) + C_{10}P_0 \int_{R_1} p_0(x)\,dx \\ &= C_{01}P_1 + \int_{R_1}\left(\underbrace{C_{10}P_0 p_0(x)}_{\geq 0} - \underbrace{C_{01}P_1 p_1(x)}_{\geq 0}\right)dx \end{aligned}$$

Risk Adjusted Classifiers

Expected Cost is then

$$E\{C(x,H)\} = C_{01}P_1 + \int_{R_1}\left(\underbrace{C_{10}P_0 p_0(x)}_{\geq 0} - \underbrace{C_{01}P_1 p_1(x)}_{\geq 0}\right)dx$$

- As in the minimum probability of error classifier, we note that all terms are positive in the integral, so to minimize expected "cost" choose R_1 to be:

$$R_1 = \{x : C_{01}P_1 p_1(x) > C_{10}P_0 p_0(x)\}$$

- Alternatively

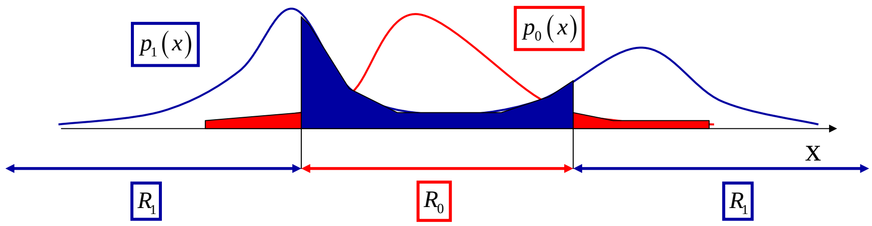
$$R_1 = \left\{x : \frac{p_1(x)}{p_0(x)} > \frac{C_{10}P_0}{C_{01}P_1}\right\}$$

- If $C_{10}=C_{01}$ then the risk adjusted classifier is equivalent to the minimum probability of error classifier.
- Another interpretation of "costs" is an adjustment to the prior probabilities.

$$\frac{P_0^{\text{adj}}}{P_1^{\text{adj}}} = \frac{C_{10}P_0}{C_{01}P_1}$$

- Then the risk adjusted classifier is equivalent to the minimum probability of error classifier with prior probabilities equal to P_1^{adj} and P_0^{adj}, respectively.

Error Probability as an Expectation

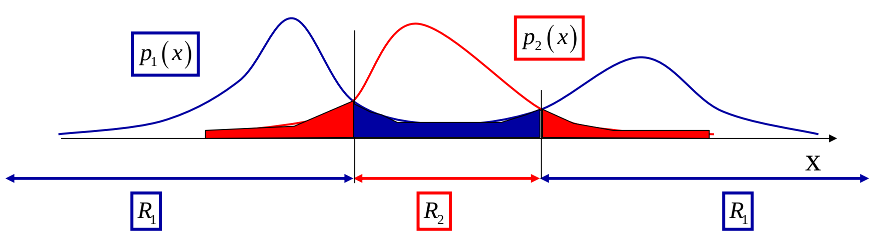


Equivalently, we can compute error probability as the expectation of a function of X and H

$$g(X,H) = \begin{cases} 1 \; ; & X \in R_1 \textbf{ AND } H_0 \text{ correct} \\ 1 \; ; & X \in R_0 \textbf{ AND } H_1 \text{ correct} \\ 0 \; ; & \text{otherwise} \end{cases}$$

$$\begin{aligned} E\{g(X,H)\} &= \int \sum_{i=1}^{2} g(x,H_i)P(H_i,x)dx \\ &= \int \sum_{i=1}^{2} g(x,H_i)P(H_i)p(x|H_i)dx \\ &= \sum_{i=1}^{2} P(H_i) \int g(x,H_i)p(x|H_i)dx \\ &= P_1 \int g(x,H_1)p_1(x)dx + P_2 \int g(x,H_2)p_2(x)dx \\ &= P_1 \left(\int_{R_1} g(x,H_1)p_1(x)dx + \int_{R_2} g(x,H_1)p_1(x)dx \right) \\ &\quad + P_2 \left(\int_{R_1} g(x,H_2)p_1(x)dx + \int_{R_2} g(x,H_2)p_1(x)dx \right) \\ &= P_1 \left(\int_{R_1} 0 \times p_1(x)dx + \int_{R_2} 1 \times p_1(x)dx \right) \\ &\quad + P_2 \left(\int_{R_1} 1 \times p_1(x)dx + \int_{R_2} 0 \times p_1(x)dx \right) \\ &= P_1 \int_{R_2} p_1(x)dx + P_2 \int_{R_1} p_2(x)dx \end{aligned}$$

Bayes Classifier vs Risk Adjusted Classifier



$$g(X,H) = \begin{cases} 1 \; ; & X \in R_1 \textbf{ AND } H_2 \text{ correct} \\ 1 \; ; & X \in R_2 \textbf{ AND } H_1 \text{ correct} \\ 0 \; ; & \text{otherwise} \end{cases}$$

$$g(X,H) = \begin{cases} C_{12} \; ; & X \in R_1 \textbf{ AND } H_2 \text{ correct} \\ C_{21} \; ; & X \in R_2 \textbf{ AND } H_1 \text{ correct} \\ 0 \; ; & \text{otherwise} \end{cases}$$

$$E\{g(X,H)\} = P_1 + \int_{R_1} \underbrace{P_2 p_2(x)}_{\text{positive}} - \underbrace{P_1 p_1(x)}_{\text{positive}} dx$$

$$E\{g(X,H)\} = C_{21}P_1 + \int_{R_1} C_{12}P_2 p_2(x) - C_{21}P_1 p_1(x)dx$$

R_1 : where $P_1 p_1(x) > P_2 p_2(x)$
R_2 : where $P_2 p_2(x) > P_1 p_1(x)$

R_1 : where $C_{21}P_1 p_1(x) > C_{12}P_2 p_2(x)$
R_2 : where $C_{12}P_2 p_2(x) > C_{21}P_1 p_1(x)$

Okay, so what.

All of this is great. We now know what to do in a few classic cases if some nice person hands us all of the probability models.

- In general we aren't given the models – What do we do?

Density estimation to the rescue.

- While we may not have the models, often we do have a collection of labeled measurements, that is a set of $\{x,H_j\}$.
- From these we can estimate the class-conditional densities. Important issues will be:
  - How "close" will the estimate be to the true model.
  - How does "closeness" impact on classification performance?
  - What types of estimators are appropriate (parametric vs. nonparametric).
  - Can we avoid density estimation and go straight to estimating the decision rule directly? (generative approaches versus discriminative approaches)