Lecture 3: Bayesian Decision Theory II

Outline:

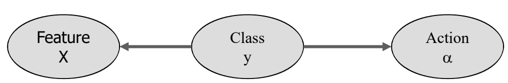
1. Bayes risk, Bayes error, and empirical error.

2. Two-state case.

3. ROC curve and PR curve.

Bayes Risk

Three key variables in Bayesian decision theory and their causal relations



We are given three functions $p(x|y)$ $p(y)$ $\lambda(\alpha|y)$

$$\alpha^* = \arg\min R(\alpha)$$

The Bayes risk is

$$\begin{aligned} R(\alpha) &= \int R(\alpha(x)|x)p(x)dx \\ &= \int \sum_y \lambda(\alpha(x)|y)p(y|x)p(x)dx \\ &= \int \sum_y \lambda(\alpha(x)|y)p(x,y)dx \end{aligned}$$

Empirical Risk

The Bayes decision theory assumes that there is underlying probability *p(x,y)*, and the Bayes risk is averaged w.r.t. to *p(x,y)*.

$$R(\alpha) = \int \sum_{y=1}^{k} \lambda(\alpha(x)|y) p(x,y) dx$$

In practice, we only have a set of labeled data (testing) drawn from *p(x,y)*,

$$D = \{(x_i, y_i): i = 1,2,\dots,m\} \sim p(x,y)$$

We can estimate the Bayes risk by accumulating all the risks i.e. wrong decisions over the testing set.

$$\hat{R}(\alpha) = \frac{1}{m}\sum_{j=1}^{m} \lambda(\alpha(x_j) \mid y_j)$$

In case of 0-1 loss: $= \frac{1}{m}\sum_{j=1}^{m} 1(\alpha(x_j) \neq y_j)$

Note the ERM paradigm: the discriminative approaches, such as Boosting, SVM minimize various upper bounds of the empirical risk.

Bayesian error

In a special case, like fish classification, the action is classification, we assume a 0/1 error.

$$\lambda(\alpha \mid y) = 0 \quad if \quad \alpha = y$$
$$\lambda(\alpha \mid y) = 1 \quad if \quad \alpha \neq y$$

The risk for classifying x to class α_i is,

$$R(\alpha = i \mid \mathrm{x}) = \sum_{\mathrm{j} \neq i} \mathrm{p}(\mathrm{y} = \mathrm{j} \mid \mathrm{x}) = 1 - \mathrm{p}(\alpha = i \mid \mathrm{x})$$

The optimal decision is to choose the class that has maximum posterior probability

$$\boxed{\alpha(\mathrm{x}) = \arg\min_{\Omega^\alpha} (1 - \mathrm{p}(\alpha \mid \mathrm{x})) = \arg\max_{\Omega^\alpha} \mathrm{p}(\alpha \mid \mathrm{x})}$$

The total risk for a decision rule, in this case, is called the Bayesian error

$$R = p(error) = \int p(error \mid \mathrm{x}) p(\mathrm{x}) d\mathrm{x} = \int (1 - p(\alpha(\mathrm{x}) \mid \mathrm{x})) p(\mathrm{x}) d\mathrm{x}$$

Bayes Risk and Bayes Error

One can minimize the Bayes risk for each x

$$
\begin{aligned}
\alpha^*(x) &= \operatorname{argmin} R(\alpha(x)|x) \\
&= \operatorname{argmin} \textstyle\sum_y \lambda(\alpha(x)|y)p(y|x) \\
&= \operatorname{argmin} 1 - p(\alpha\,|x) \\
&= \operatorname{argmax} p(\alpha\,|\,x)
\end{aligned}
$$

That is, you always choose the class that is most probable.

Bayes error is the minimum error (lower bound).

What if we make randomize decisions, i.e. proportional to the posterior probability?

$$\alpha_{rand}(x) \sim p(y|x)$$

The loss will be bigger than the Bayesian decision

$$R(\alpha_{rand}(x)) = 1 - \sum_y p^2(y|x)$$

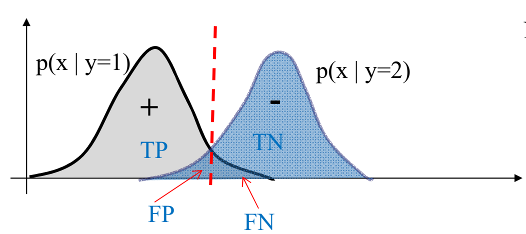
Two-State Case

Detect “target” or “non-target” (say human “face” or “non-face”).
The decision boundary will be decided by the equation with rish function being a 2x2 matrix

λ_{11}	λ_{12}
λ_{21}	λ_{22}

$$R(\alpha = 1|x) = R(\alpha = 2|x)$$

$$\frac{p(x|y=1)}{p(x|y=2)} = \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{p(y=2)}{p(y=1)} = T$$



Note:
hit rate (True Positive)
missing rate (False Negative)

correct rejection (True Negative)
false alarm (False Positive)

ROC curves

For two-state problems, the Bayes decision rule is where T depends on the priors and the loss function.

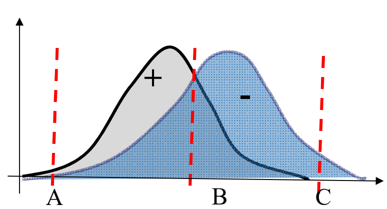
$$\log \frac{p(x|y=1)}{p(x|y=2)} > T$$

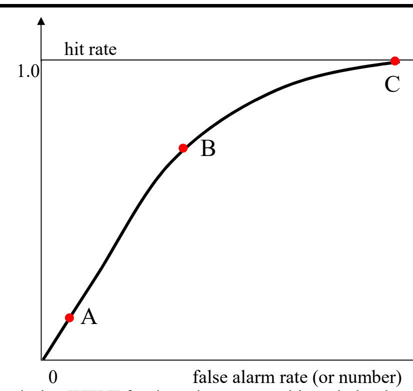
A Receiver Operator Characteristics (ROC) curve plots the proportion of correct responses (hits) against the false positives as the threshold T changes.

It is more general than Bayes risk as it is independent of the observer's loss function.

ROC curve

Moving the decision threshold T from left to right, we obtain a (hit-rate, false-alarm) at each T, and thus plot a curve





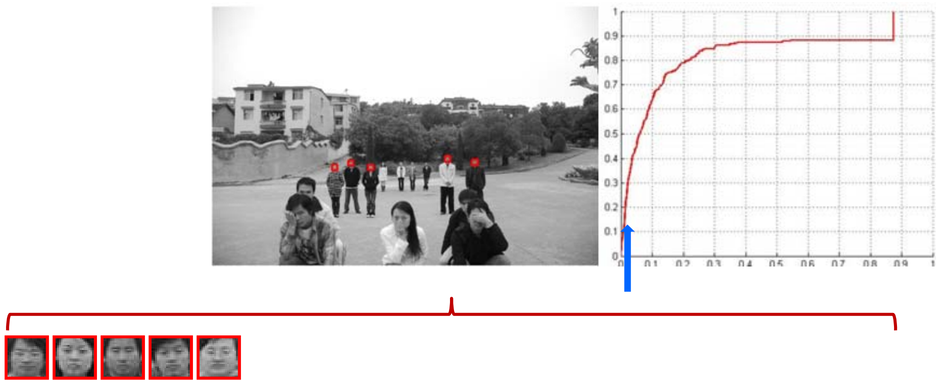
Remarks:

1) The ROC curve was used by radar engineers during WW II for detecting enemy objects in battle fields, also known as the signal detection theory. It was later used in psychology for signal detection (say a dim light) in perception.

2) The 2 distributions plotted above are $p(x|y = +1)$ vs. $p(x|y= -1)$. The prior probability (i.e. the population sizes) for the +/- classes is not factored in. To count for the prior probability, we can plot $p(y=+1|x)$ vs $p(y=-1 |x)$ for every x. When x is in hi-dimensional space, then the threshold equation $p(y=+1|x) / p(y=-1 |x) = T$ corresponds to a moving boundary in the space of x.

Example: detecting faces

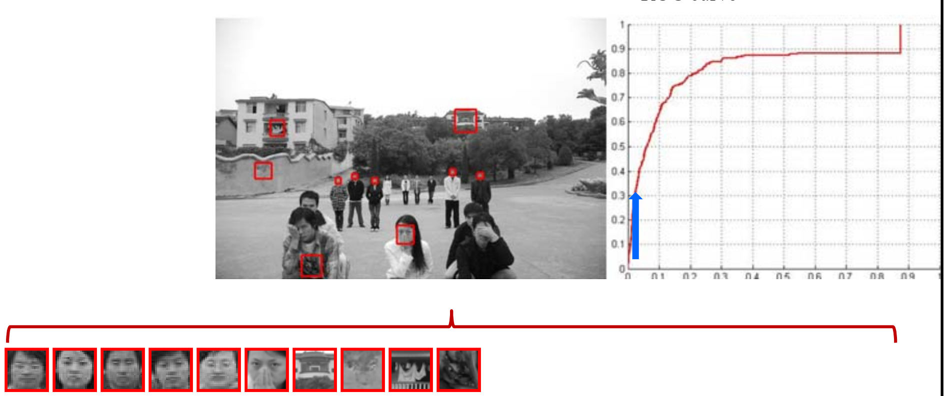
ROC curve



Example from Tianfu Wu 2009

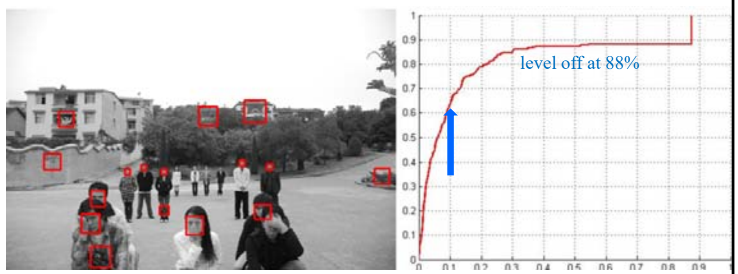
Example: detecting faces

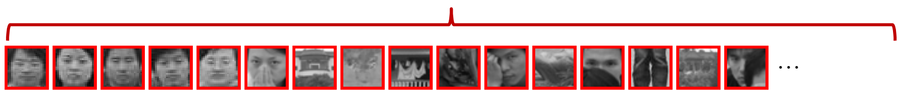
ROC curve



Example: detecting faces

ROC curve

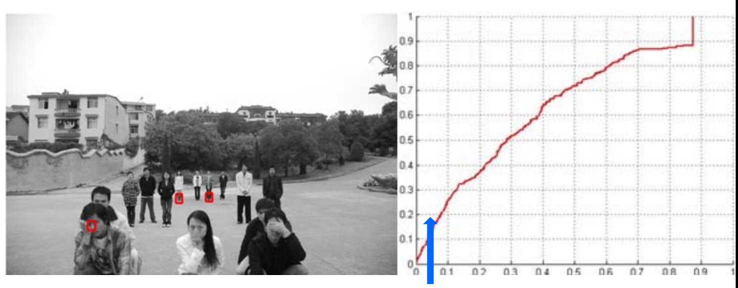


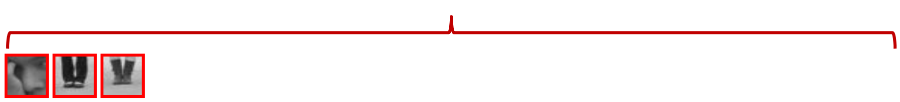


…

Example: detecting noses

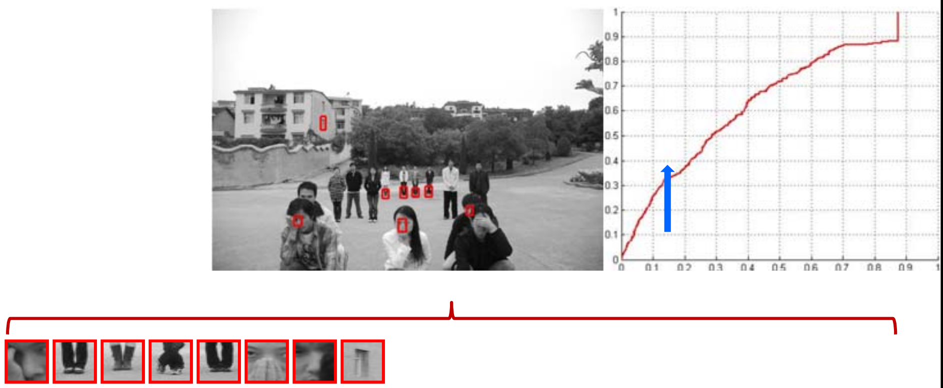
This ROC curve is much lower, usually the area under curve (**AUC**) measures the effectiveness.





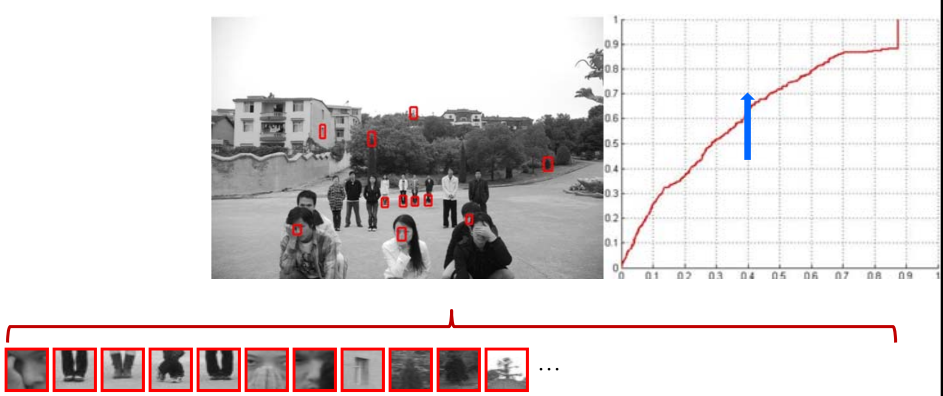
Example: detecting noses

ROC curve



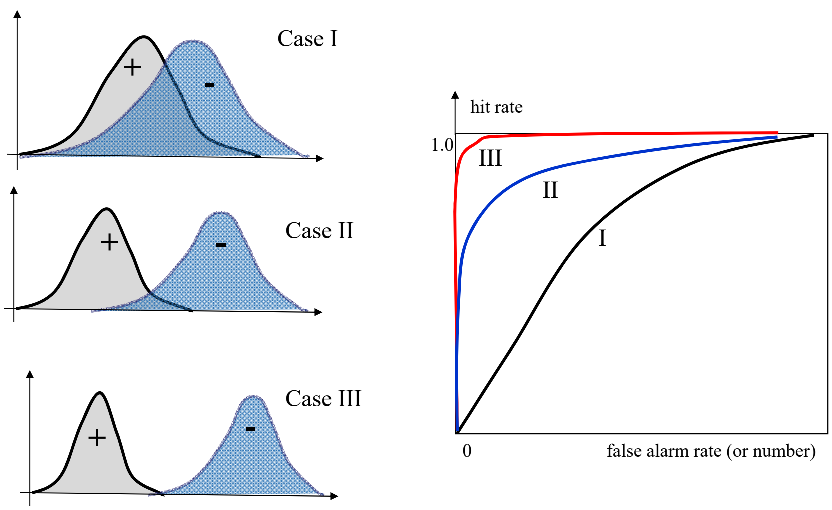
Example: detecting noses

ROC curve



After this, it is almost random, you still get a 100% detection if we cover all windows in the image (huge number of false alarms).

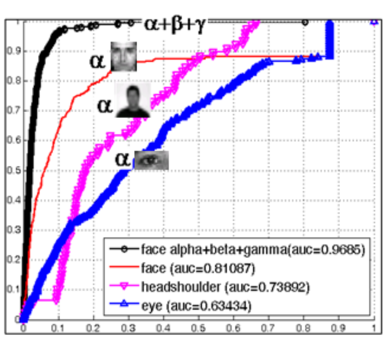
ROC curves



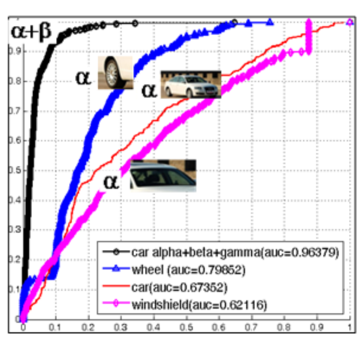
ROC

The ROC curves are independent of threshold, but dependent of features and models. But there should be an upper limit regardless of features and models.

In vision, the effectiveness for objects (classes) quite drastically.



(a) head/shoulder---face---eyes



(b) car---parts

Another Example: Edge Detection

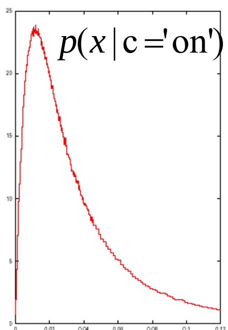
The boundaries of objects (right) usually occur where the *image intensity gradient* is large (left).

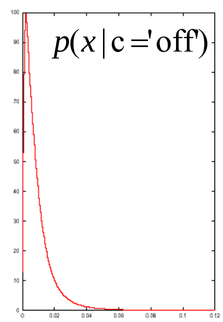
Lecture notes Stat 231-CS276A, © S.C. Zhu

Example: Boundary Detection

Learn the probability distributions for intensity gradient at position v being *on* and *off* labeled edges. We have two classes:

c in {'on' , 'off'} $x = |\nabla I|$

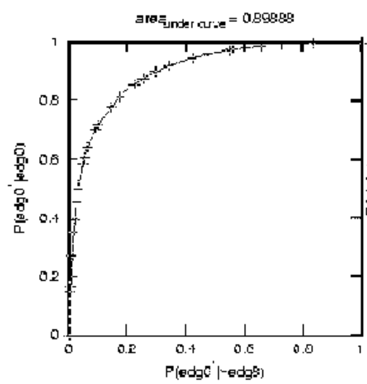


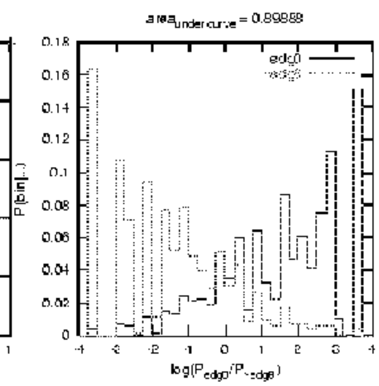


Lecture notes Stat 231-CS276A, © S.C. Zhu

Example on boundary detection

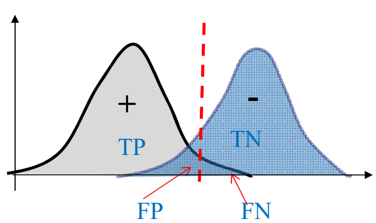
Perform edge detection by log-likelihood ratio test.





Precision-Recall Curves

In many applications, such as object detection or information retrieval (search engine), we have target (+) and background (-) classes, it is hard to define the volume of the negatives (background is almost infinite). We only care about the target class (TP, FP).



$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

$$\text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

Precision is a measure of correctness
--- the percentage of detected objects that are true.

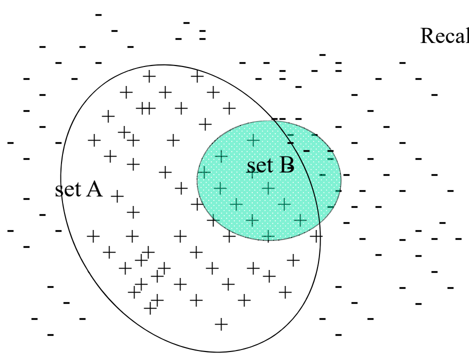
Recall is a measure of completeness
--- the percentage of true objects that are detected.

Precision-Recall Curves

For example, in retrieval, suppose the total relevant pages are in set A, a search engine returns a set B of pages (see the circle below)

$$\text{Precision} = \frac{tp}{tp + fp}$$ is the percentage in set B

$$\text{Recall} = \frac{tp}{tp + fn}$$ is the percentage in set A

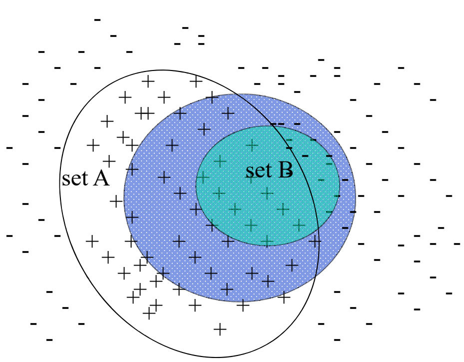


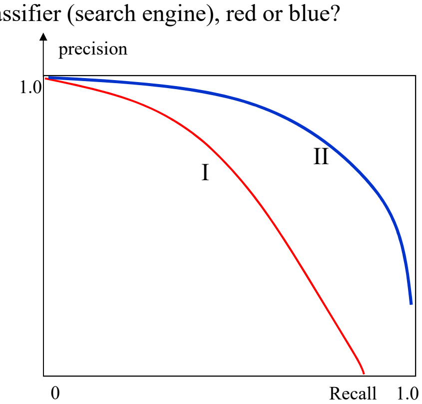
When do we use PR not ROC?

When the negative is massive and the false negative rate can be made arbitrarily small by adding easy (irrelevant) negatives to alter the negative population, and thus the ROC curve is no longer meaningful.

Precision-Recall Curves

When we enlarge the circle (lower the search criterion), we get a PR-curve
Different sequences of set B's lead to different PR-curves.
In the right figure which one is a better classifier (search engine), red or blue?





The area under the PR curve is usually used as a single number performance measure.

