Lecture 2: Bayesian Decision Theory I

Bayesian decision theory is the basic framework for pattern recognition.

Outline:

1. Diagram and formulation
2. Bayes rule for inference
3. Bayesian decision
4. Discriminant functions and space partition
5. Advanced issues

Diagram of pattern classification

Procedure of pattern recognition and decision making

X--- all the observables using existing sensors and instruments
x --- is a set of features selected from components of X or linear/non-linear functions of X.
which can be hand-crafted or learned from raw data directly
p(y | x) --- is our belief/perception about the subject class with uncertainty represented by probability.
α --- is the action or decision that we take for x.

We denote the three spaces (sets) by

$$x \in \Omega^d, \quad y \in \Omega^C, \quad \alpha \in \Omega^\alpha$$

$$x = (x_1, x_2, ..., x_d) \text{ is a vector}$$

$$y \text{ is the index of classes, } \Omega^C = \{1, 2, ..., k\}$$

Examples

Ex 1: Fish classification

X = I is the image of fish,

x =(brightness, length, #fin, ….)

y is our belief what the fish type is
Ω^c={"sea bass", "salmon", "trout", …}

α is a decision for the fish type,
in this case $\Omega^c = \Omega^\alpha$
Ω^α ={"sea bass", "salmon", "trout", …}

Ex 2: Medical diagnosis

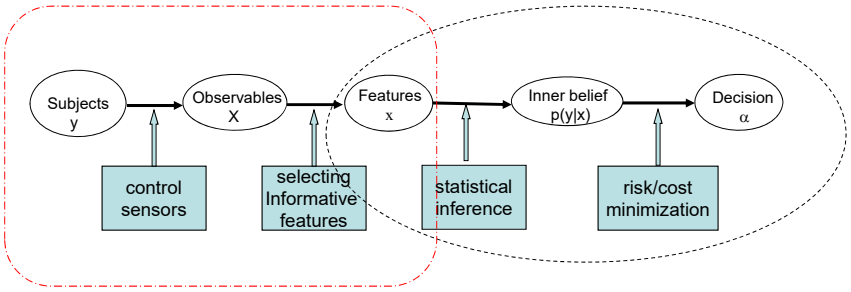
X= all the available medical tests, imaging scans (ultra sound, CT, blood test) that a doctor can order for a patient
x =(blood pressure, glucose level, …, shape)

y is an illness type
Ω^c={"cold", "TB", "pneumonia", "lung cancer",…}

α is a decision for treatment,
Ω^α ={"Tylenol", "Hospitalize", …}

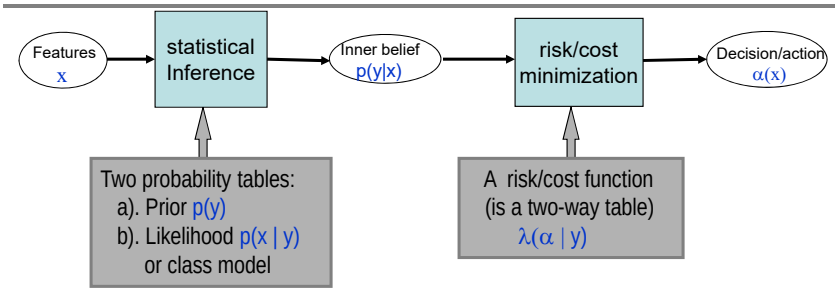
Note that it only outputs a class label, later, the output can be a structured description (i.e. parse tree).

Tasks



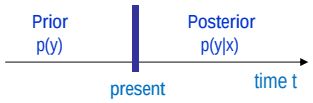
In Bayesian decision theory, we are concerned with the last three steps in the big ellipse assuming that the observables are given and features are selected. This part is automated following standard code and procedure in *Machine Learning*.
The problems are in the rectangular box and *domain specific*: to select effective features.
It is unclear i) why we extract features, ii) whether we increase or reduce dimensions, and iii) why we develop and compute inner belief before decision.

Bayesian Decision Theory



The belief or probability on the class y is computed by the Bayes rule

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}$$



The expected risk is computed by

$$R(\alpha_i \mid x) = \sum_{j=1}^{k} \lambda(\alpha_i \mid y = j)p(y = j \mid x)$$

Recall the fish classification example

Salmon

Sea Bass

- Two-class case: any given fish is either Salmon or Sea Bass (i.e. state of nature of the fish) ---- or fish v.s. non-fish
- State of nature is a random variable y, with prior probability P(y) – reflects our knowledge / belief of how likely we expect a certain fish before actually observing the data
  - $Y = y_1$ for Sea Bass
  - $Y = y_2$ for Salmon
- Decision rule with only the prior information
  - Decide y_1 if $P(y_1) > P(y_2)$ otherwise decide y_2 ==> Issues?

A simple example

Suppose you are betting on which side a tossed coin will fall on. You win $1 if you are right and lose $1 if you are wrong.
Let x be whatever evidence that we observe. Together with some prior information about the coin, you believe the head has a 60% chance and tail 40%.

$$y \in \{H, T\}, \quad \alpha \in \{H, T\}, \quad \lambda(\alpha(x) \mid y) = \begin{matrix} & y \\ \alpha & \begin{matrix} -1, & +1 \\ +1, & -1 \end{matrix} \end{matrix}$$

$$R(\alpha_i \mid x) = \sum_{j=1}^{k} \lambda(\alpha_i \mid y = j) p(y = j \mid x)$$

$$= \begin{matrix} & y \\ \alpha & \begin{matrix} -1, & +1 \\ +1, & -1 \end{matrix} \end{matrix} \begin{matrix} y \\ \begin{matrix} 0.6 \\ 0.4 \end{matrix} \end{matrix}$$

$$= \alpha \begin{matrix} -0.2 \\ 0.2 \end{matrix}$$

Here risk= -1 means reward 1.

So, to minimize the risk, we choose head α = H.

*** α and y are shown as indices to the rows/columns of the matrix.*

Decision Rule

A decision rule is a mapping function from feature space to the set of actions

$$\alpha(x): \Omega^d \to \Omega^\alpha$$

we will show that randomized decisions won't be optimal.

A decision is made to minimize the average (expected) cost (or risk),

$$R = \int R(\alpha(x) \mid x) p(x) \, dx$$

It is minimized when our decision is made to minimize the cost (or risk) for each instance x.

$$\alpha(x) = \arg\min_{\Omega^\alpha} R(\alpha \mid x)$$

$$= \arg\min_{\Omega^\alpha} \sum_{j=1}^{k} \lambda(\alpha \mid y = j) p(y = j \mid x)$$

$$= \arg\min_{\Omega^\alpha} \sum_{j=1}^{k} \lambda(\alpha \mid y = j) p(x \mid y = j) p(y = j)$$

** When we replace the expectation by the average risk of a set of training sample, it is called the *empirical risk*

$$\hat{R} = \frac{1}{m} \sum_{i=1}^{m} R(\alpha(\alpha_i) \mid x_i), \_ \text{with} \_ \{x_i\} \sim p(x)$$

Special case: 0-1 loss

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

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

In this case, the risk for classifying x to class α=i is the *probability of mis-classification*

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

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

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

Discriminant functions

To summarize, we take an action to maximize some discriminant functions:

$$\mathrm{g_i(x)} = \mathrm{p(y = i \mid x)}$$
$$\mathrm{g_i(x)} = \mathrm{p(x \mid y = i)p(y = i)}$$
$$\mathrm{g_i(x)} = \log \mathrm{p(x \mid y = i)} + \log \mathrm{p(y = i)}$$
$$\mathrm{g_i(x)} = -\mathrm{R}(\alpha = \mathrm{i} \mid \mathrm{x})$$ i.e. Bayes or minimum conditional risk discriminant

$$\alpha(\mathrm{x}) = \mathrm{argmax}\ \{\mathrm{g_1(x)}, \mathrm{g_2(x)}, \ldots\ldots, \mathrm{g_k(x)}\}$$

Two-class Discriminants

Consider a single discriminant function

$$g(x)=g_1(x)-g_2(x)$$

A simple rule: decide y_1 if g(x) > 0 otherwise decide y_2

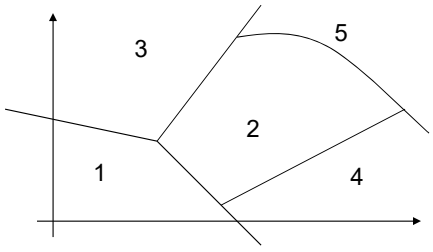
e.g., $g(x)=p(y_1|x)-p(y_2|x)$

Partition of feature space

$$\alpha(x):\ \Omega^{d} \rightarrow \Omega^{\alpha}$$

The decision is a partition or coloring of the feature space into k subspaces.

$$\Omega=\cup_{i=1}^{k}\Omega_i \qquad \Omega_i \cap \Omega_j = \phi,\ i \neq j$$



The decision boundaries between two classes i and j are decided by the equation.

$$g_i(x) = g_j(x) \qquad \forall\ i \neq j$$

Discussion: other decision rule

With 0-1 loss function, the Bayes decision chooses the class with highest posterior probability.

$$\alpha(\mathrm{x})_{\mathrm{Bayes}} = \arg\min_{\alpha\in\Omega^{\alpha}} (1-\mathrm{p}(\alpha\,|\,\mathrm{x})) = \arg\max_{\alpha\in\Omega^{\alpha}} \mathrm{p}(\alpha\,|\,\mathrm{x})$$

The Bayes decision is said to yield the minimum error among all possible decision rules.

Discussion:

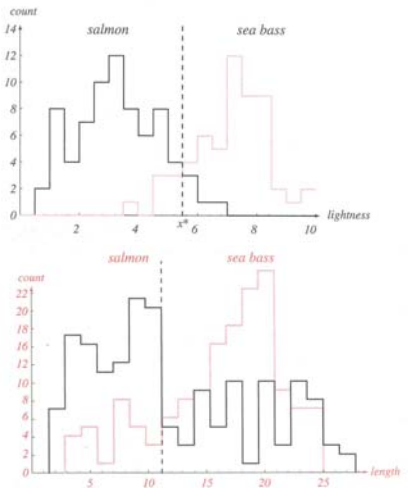
1, What are the assumptions of the conclusion above?

2, One may argue for a randomized decision, will it lead to better results?

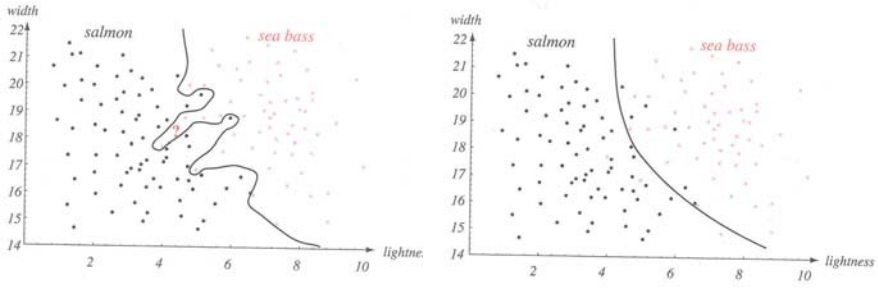
$$\alpha(\mathrm{x})_{Random} \sim \mathrm{p}(\alpha\,|\,\mathrm{x})$$

The decision will be proportional to the probability.

An example of fish classification



Decision/classification Boundaries



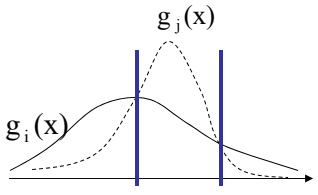
Decision boundaries:

The left boundary has zero training error – that is often a desirable goal in machine learning. but it is too specific to the data, and is often said to be "over-fitting". If we draw a different sample from the data, the boundary will be very different. Such boundary leads to higher error on testing data.

Decision boundaries

The decision boundaries between two classes may have multiple pieces and the domain Ω_i of a class may consist of many components.

$$g_i(x) = g_j(x) \qquad \forall\, i \neq j$$



Example: analytic solution of the decision boundary

For simplicity, people often assume Gaussian probabilities for the class models.

$$p(x|w_i) = \frac{1}{(2\pi)^{d/2}\det(\Sigma_i)^{1/2}} \exp\{-\frac{1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i)\}.$$

We consider a case when $\Sigma_i = \sigma^2 I$

Then the discriminant function is

$$g_i(x) = -\frac{1}{2\sigma^2}[x^t x - 2\mu_i^t x + \mu_i^t \mu_i] + \ln p(w_i)$$

The boundary equation is a straight line --- linear discriminant or linear machine

$$(\mu_i - \mu_j)^t (x - x_0) = 0$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{|\mu_i - \mu_j|^2} \ln \frac{p(w_i)}{p(w_j)} (\mu_i - \mu_j)$$

Close-form solutions

In two case classification, k=2, with p(x|w) being Normal densities, the decision boundaries can be computed in close-form.

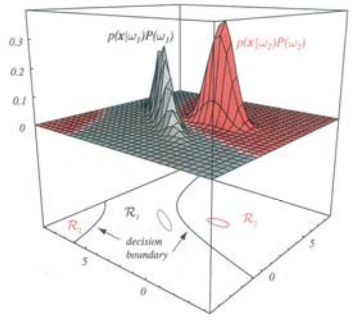


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution.

Other cases:

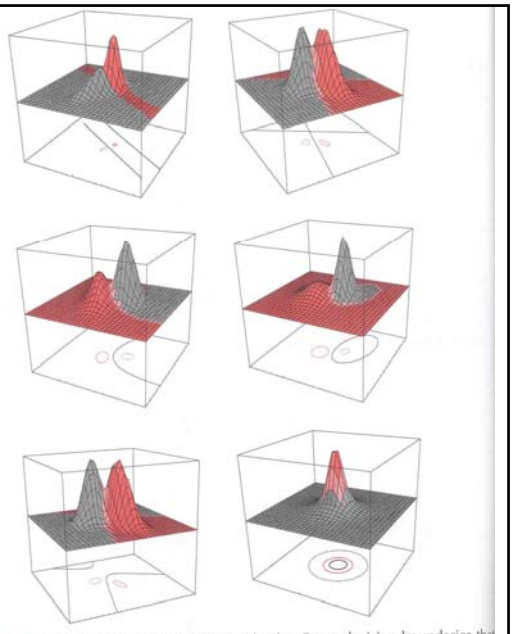


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability density.

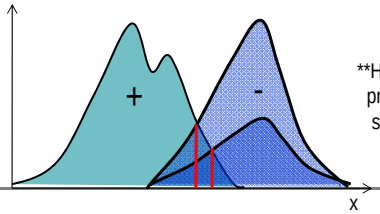
Discussion of advanced issues

1. "Subjectivism" of the prior in Bayesian decision

Some people may accept that risk/cost can affect the decision boundary, but don't like the fact that the prior probability (i.e. the population or frequency of a certain class) affect the decision boundary, as is shown in the figure below.

This means that the Bayesian decision on a certain input "x" is not entirely based on x *individually*, but also accounts for the *sub-population* of the class collectively. So Bayesian decision is a "stereotypical" decision to minimize overall risk/cost, and doing so may cause *higher rate* of mis-classification to some individuals.

Civil rights prevent us from using certain features (such as gender, race) in computing risk or taking actions, though doing so may be more cost effective.



**Here is an example that the prior probability about the negative population shifts the decision boundary.

Discussion of advanced issues

2: Learning the probabilities p(y), p(x|y) or p(y|x) --- the subject of machine learning

There are two important factors to consider:

(i) Generative vs. discriminative: which is more effective?
This depends on the structures of the feature space.

(ii) Should we learn the probability or just the probability ratio?

$$\frac{p(y=-1\,|\,x)}{p(y=+1\,|\,x)}$$

This will choose different features which are most discriminative.

How many examples are enough for learning a model/concept?
--- learnability i.e. the probably and approximately correct (PAC)-learning.

Discussion of advanced issues

3, Context or sequential information in classification.

So far, the classification is based on individual input x based on a given model (which is assumed to be learned off-line). In general, one only have very scarce data, therefore we need to consider

(i) Recursive learning and online adaptation of the model

---- This is particularly useful for object tracking.

(ii) Active learning and Markov decision process
exploring new features based on current results.