8001652 Introduction to Pattern Recognition. Lectures 4 and 5: Bayesian decision theory

Jussi Tohka

jussi.tohka@tut.fi

Institute of Signal Processing

Tampere University of Technology

Bayesian decision theory

- Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification.
- This approach is based on quantifying the tradeoffs between various classification decisions and their costs.
- The problem needs to be posed in probabilistic terms and associated probabilities need all to be completely known.
- The theory is just a formalization of some common sense procedures.

The fish example: Terminology

- Separate between two kinds of fish: Sea bass and salmon. No other kinds of fish are possible.
- ω is the state of the nature, it is a random variable.
- Two states are possible: $\omega = \omega_1$ for sea-bass and $\omega = \omega_2$ for salmon.
- If the sea contains more sea-basses than salmons, it is natural to assume (even when no data or features are available), that the caught fish is a sea-bass. This is modeled with the prior probabilities, $P(\omega_1)$ and $P(\omega_2)$, which are positive and sum to one.
- If there are more sea basses, $P(\omega_1) > P(\omega_2)$.

The fish example continued

- If we must decide at this point (for some curious reason) which fish we have, how would we decide?
- Because there are more sea-basses , we would say that the fish is a sea-bass.
- In other words, our decision rule becomes: Decide ω_1 if $P(\omega_1) > P(\omega_2)$ and otherwise decide ω_2.
- To develop better rules, we must extract some information or features from the data.
- This means, for instance, making lightness measurements about the fish.

The fish example continued

- Ok, suppose we have a lightness reading, say x, from the fish. What to do next?
- We know every probability relevant to the classification problem. That is we know $P(\omega_1), P(\omega_2)$ and *the class conditional probability densities* $p(x|\omega_1)$ and $p(x|\omega_2)$.
- Based on these we can compute the probability that the state of the nature $\omega = \omega_1$ given that the lightness reading is x and similarly for salmon. Just use Bayes formula:

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

where the *evidence* $p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j)$. The evidence is merely a scale factor. We do not have to compute it at this point.

The fish example continued

- The decision rule: Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$ and otherwise decide ω_2.
- Remember that we do not know (directly) $P(\omega_j|x)$. They must be computed through the Bayes rule:

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

The fish example continued

- The justification for the rule:
- $P(error|x) = P(\omega_1|x)$ if we decide ω_2.
  $P(error|x) = P(\omega_2|x)$ if we decide ω_1.
- Average error

$$P(error) = \int P(error, x)dx = \int P(error|x)p(x)dx.$$

- Thus, if $P(error|x)$ is minimal for every x, also the average error is minimized.
- The decision rule 'Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$ and otherwise decide ω_2.' guarantees just that.

The fish example continued

- An equivalent decision rule is obtained by multiplying $P(\omega_j|x)$ in the previous rule by $p(x)$. Because $p(x)$ is a constant this obviously does not affect the decision itself.
- We have a rule:
  Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$, otherwise decide ω_2.

BDT - terminology

- In BDT, the sample space is often called the feature space i.e. the space to which the features extracted belong.
- For moment, assume that the feature space is $\mathbb{R}^d$.
- A *loss function* tells how costly each action is and is used to convert a probability determination into a decision.
- Write $\omega_1, \ldots, \omega_c$ for the c states of nature, or categories, or classes.
- Write $\alpha_1, \ldots, \alpha_a$ for a possible actions.
- The loss function $\lambda(\alpha_i|\omega_j)$ describes the loss incurred for taking action α_i if the state of nature ω_j.

BDT - terminology

- If the true state of nature is ω_j, by definition we will incur the loss $\lambda(\alpha_i|\omega_j)$ when taking the action α_i after observing $\mathbf{x}$.
- The conditional risk of taking action α_i, after observing $\mathbf{x}$, is

$$R(\alpha_i|\mathbf{x}) = \sum_{i=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x}).$$

- Denote the action taken after observing $\mathbf{x}$ by $\alpha(\mathbf{x})$. The (total) expected loss, termed overall risk, is

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x}) p(\mathbf{x}).$$

- Function $\alpha(\mathbf{x})$ is called the decision rule.

Bayes decision rule

- Now, we would like to derive such decision rule $\alpha(\mathbf{x})$ that it minimizes the overall risk.
- It is not hard to guess that this decision rule is

  Select action α_i that gives the minimum expected loss $R(\alpha_i|\mathbf{x})$.
- This is called the *Bayes decision rule*.
- The resulting minimum overall risk achieved by taking actions according to Bayes decision rule is called *Bayes risk*.

Two category classification

- The possible states of nature are ω_1, ω_2. The action α_1 corresponds deciding that the true state of nature is ω_1 and α_2 corresponds deciding that the true state of nature is ω_2.
- $\lambda(\alpha_i|\omega_j) = \lambda_{ij}$
- $$\begin{aligned} R(\alpha_1|\mathbf{x}) &= \lambda_{11} P(\omega_1|\mathbf{x}) + \lambda_{12} P(\omega_2|\mathbf{x}) \\ R(\alpha_2|\mathbf{x}) &= \lambda_{21} P(\omega_1|\mathbf{x}) + \lambda_{22} P(\omega_2|\mathbf{x}) \end{aligned}$$
- The fundamental rule is to decide that the true state of nature is ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$ and ω_2 otherwise.

Two category classification

- We decide (that the true state of nature is) ω_1 if

$$(\lambda_{21} - \lambda_{11})P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2|\mathbf{x}).$$

- Ordinarily $\lambda_{21} - \lambda_{11}$ and $\lambda_{12} - \lambda_{22}$ are positive. i.e. The loss is greater when making a mistake.
- Assume $\lambda_{21} > \lambda_{11}$. Then an equivalent rule is: Decide ω_1 if

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}.$$

Zero-one-loss

- Particularly interesting loss function is zero-one-loss

$$\lambda(\alpha_i|\omega_j) = \begin{matrix} 0 & \text{if} & i = j \\ 1 & \text{if} & i \neq j \end{matrix}$$

- The matrix form for the two category case

$$\begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}$$

- This loss function leads to minimum error rate classification:

  Decide ω_i if $P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x})$ for all $j \neq i$.

Recap: The Bayes Classifier

- Given a feature vector $\mathbf{x}$, compute the conditional risk for taking action α_i for all $i = 1, \ldots, a$ and select the action that gives the **smallest** conditional risk $R(\alpha_i|\mathbf{x})$.
- Classification with zero-one-loss: Compute the probability $P(\omega_i|\mathbf{x})$ for all categories $\omega_1, \ldots, \omega_c$ and select the category that gives the **largest** probability. Remember to use Bayes rule in computing the probabilities.

Discriminant functions

- Classifiers can be represented in terms of discriminant functions $g_1(\mathbf{x}), \ldots, g_c(\mathbf{x})$ for c category classifier.
- The classifier then assigns a feature vector $\mathbf{x}$ to class ω_i if

$$g_i(\mathbf{x}) > g_j(\mathbf{x})$$

  for all $i \neq j$.
- For the Bayes classifier $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$.

Discriminant functions

- A classifier can be represented by a different sets of discriminant functions. In other words, the choice of discriminant functions is not unique.
- Example: Equivalent discriminant functions for the minimum error rate classifier
  - $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)}$
  - $g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$
  - $g_i(\mathbf{x}) = \log p(\mathbf{x}|\omega_i) + \log P(\omega_i)$
- We prefer the simplest discriminant functions for a particular classifier.

Discriminant functions

- Two-category case: We may combine the two discriminant functions into a single discriminant function.
- The decision rule: Decide ω_1 if $g_1(\mathbf{x}) > g_2(\mathbf{x})$ and otherwise decide ω_2.
- Define $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$.
- We obtain an equivalent decision rule: Decide ω_1 if $g(\mathbf{x}) > 0$.

Discriminant functions

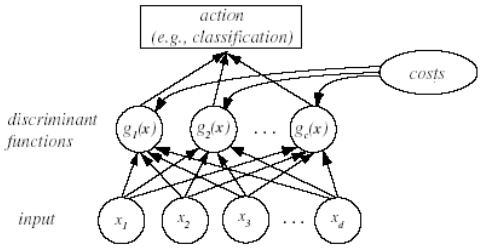

Figure 2.5 from Duda, Hart, Stork: Pattern Classification, Wiley, 2001

Decision regions

- The effect of any decision rule is to divide the feature space (in this case $\mathbb{R}^d$) into c *decision regions*, $\mathcal{R}_1, \ldots, \mathcal{R}_c$.
- Decision rules can be written as if $\mathbf{x} \in \mathcal{R}_i$ decide ω_i.
- Decision regions from discriminant functions:

$$\mathcal{R}_i = \{\mathbf{x} : g_i(\mathbf{x}) > g_j(\mathbf{x}) \forall i \neq j\}.$$

- Boundaries of decision regions, i.e places where two or more discriminant functions yield the same value, are called decision boundaries.

Decision regions

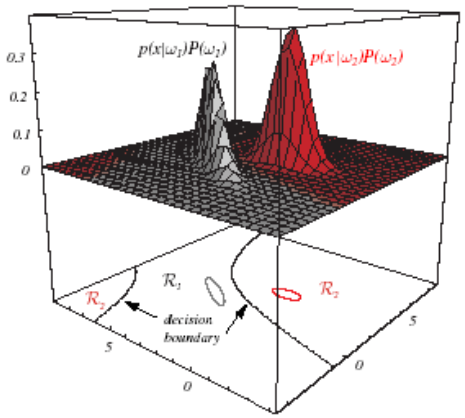

Figure 2.6 from Duda, Hart, Stork: Pattern Classification, Wiley, 2001

The normal density

- Univariate case

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2],$$

  where parameter $\sigma > 0$.
- Multivariate case

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \exp[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)],$$

  where $\mathbf{x}$ is a d-column vector and Σ is a positive definite matrix.
- For a positive definite matrix Σ, $\mathbf{x}^T\Sigma\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$.

The normal density - properties

- The normal density has several properties which give it a special position among probability densities. To large extent, this is due analytical tractability - as we shall soon see - but there are also other reasons for favoring normal densities.
- $X \sim N(\mu, \Sigma)$ stands for 'X is a RV having the normal density with parameters μ, Σ.'
- The expected value of X $E[X] = \mu$.
- The variance of X $Var[X] = \Sigma$.

The normal density - properties

- Let $X = [X_1, \ldots, X_d] \sim N(\mu, \Sigma)$. Then $X_i \sim N(\mu_i, \sigma_{ii})$. Let A, B be $d \times d$ matrices. Then $AX \sim N(A\mu, A\Sigma A^T)$. AX and BX are independent if and only if $A\Sigma B^T$ is the zero matrix.
- The sum of two normally distributed RVs is also normally distributed.
- Central Limit Theorem: The sum of n identically distributed independent (i.i.d) RVs tends to a normally distributed RV as n approaches infinity.

DFs for the normal density

- The minimum error rate can be achieved by use of DFs

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i).$$

- Letting $p(\mathbf{x}|\omega_i) = N(\mu_i, \Sigma_i)$ we obtain

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\det(\Sigma_i) + \ln P(\omega_i).$$

DFs for the normal density: $\Sigma_i = \sigma^2 I$

- $\Sigma_i = \sigma^2 I$: Features are independent and each feature has the same variance σ^2.
- Geometrically, this corresponds to the situation in which the samples fall in equal-size (hyper)spherical clusters, the cluster for the ith class being centered around μ_i.

DFs for the normal density: $\Sigma_i = \sigma^2 I$

- Let's now simplify the DFs

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\det(\Sigma_i) + \ln P(\omega_i).$$

- 1) All the constant terms like $\frac{d}{2}\ln 2\pi$ can be dropped - dropping them does not affect the classification result.
- 2) In this particular case also the determinants of the covariance matrices have all the same value (σ^{2d}).
- 3) $\Sigma^{-1} = \frac{1}{\sigma^2}I$ and hence

$$g_i(\mathbf{x}) = -\frac{||\mathbf{x}-\mu_i||^2}{2\sigma^2} + \ln P(\omega_i).$$

DFs for the normal density: $\Sigma_i = \sigma^2 I$

- (1)

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{||\mathbf{x}-\mu_i||^2}{2\sigma^2} + \ln P(\omega_i) \\ &= -\frac{1}{2\sigma^2}(\mathbf{x}-\mu_i)^T(\mathbf{x}-\mu_i) + \ln P(\omega_i) \\ &= -\frac{1}{2\sigma^2}(\mathbf{x}^T\mathbf{x} - 2\mu_i^T\mathbf{x} + \mu_i^T\mu_i) + \ln P(\omega_i). \end{aligned}$$

- From the last expression we see that the quadratic term $\mathbf{x}^T\mathbf{x}$ is same for all categories and can be dropped. Hence, we obtain equivalent linear discriminant functions:

$$g_i(\mathbf{x}) = \frac{1}{\sigma^2}(\mu_i^T\mathbf{x} - \frac{1}{2}\mu_i^T\mu_i) + \ln P(\omega_i).$$

Linear discriminant functions

- Discriminant functions that can be written as

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

  are said to be linear.
- A classifier that uses linear decision functions is called linear machine.
- w_{i0} is called the *threshold* or *bias* for ith category.
- Decision boundaries are hyperplanes.

Minimum distance classifier

- If we assume that all $P(\omega_i)$ are equal and $\Sigma_i = \sigma^2 I$, we obtain the minimum distance classifier.
- Note that this means that $P(\omega_i) = \frac{1}{c}$.
- The name of the classifier follows from the set of discriminant functions used:

$$g_i(\mathbf{x}) = -||\mathbf{x} - \mu_i||.$$

- Hence, a feature vector is assigned to the category with the nearest mean.
- Note that a minimum distance classifier can be also implemented as a linear machine.

DFs for the normal density: $\Sigma_i = \Sigma$

- Consider a bit more complicated model. Now all covariance matrices have still the same value but there exists some correlation between the features. In other words, features are not independent anymore.
- Also in this case we have a linear machine:

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0},$$

  where

$$\mathbf{w}_i = \Sigma^{-1} \mu_i$$

  and

$$w_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \ln P(\omega_i).$$

DFs for the normal density: Σ_i arbitrary

- It is time to consider the most general Gaussian model, where features from each category are assumed to be normally distributed, but nothing more is not assumed.
- In this case the discriminant functions

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1} (\mathbf{x}-\mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln \det(\Sigma_i) + \ln P(\omega_i).$$

  cannot be simplified, except for dropping the constant terms.
- Discriminant functions are now necessarily quadratic which means that decision regions may have more complicated shapes than in the linear case.

BDT - Discrete features

- In many practical applications, the feature space is discrete. The components of the feature vectors are then binary or higher-integer valued.
- This simply means that integrals as in continuous case must be replaced with sums and probability densities must be replaced with probabilities.
- The minimum error rate classification is then: Decide ω_i if

$$P(\omega_i|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})} > P(\omega_j|\mathbf{x}).$$

Independent binary features

- Consider the two-category problem, where the feature vectors $\mathbf{x} = [x_1, \ldots, x_d]$ are binary, i.e. x_i is either 0 or 1.
- Assume further that features are (conditionally) independent, that is

$$P(\mathbf{x}|\omega_j) = \prod_{i=1}^{d} P(x_i|\omega_j).$$

- Denote $p_i = P(x_i = 1|\omega_1)$ and $q_i = P(x_i = 1|\omega_2)$.

Independent binary features

- Then,
- $P(\mathbf{x}|\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}$
- $P(\mathbf{x}|\omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}$
- $g(\mathbf{x}) = \sum_{i=1}^{d}[x_i \ln \frac{p_i}{q_i} + (1-x_i)\ln \frac{1-p_i}{1-q_i}] + \ln \frac{P(\omega_1)}{P(\omega_2)}$
- $g(\mathbf{x}) = \sum_{i=1}^{d} x_i \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} + \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$

Independent binary features

- Note that we have a linear machine: $g(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + w_{i0}$.
- The magnitude of w_i determines the importance of the 'yes' (1) answer for x_i. If $p_i = q_i$, the value of x_i gives no information about the state of nature.
- The prior probabilities appear only in the bias term. Increasing $P(\omega_1)$ biases the decision in favor of ω_1.

Receiver operating characteristic

- For signal detection theory and receiver operating characteristic (ROC), see www.cs.pitt.edu/~milos/courses/cs2750/lectures/class9.pdf .

BDT - context

- This far we have assumed that our interest is in classifying a single object at time.
- However, in applications we may need to classify several objects at same time. Example: image segmentation.
- If we assume that the state of nature of one object is independent from remaining ones, nothing does change.
- If there is some dependence, then the basic principles remain the same, i.e. we assign objects to the most probable category. But now we need to take also categories of other objects into account and place all objects in such categories that the probability of the whole ensemble is maximized.
- Computational difficulties follow.