Bayesian decision theory

R.O. Duda, P.E. Hart, D.G. Stork

Pattern Classification - Chapter 2

Introduction

- **Bayesian decision theory** = fundamental statistical approach to the problem of pattern classification, based on quantifying the tradeoffs between various classification decisions using probability and the costs that accompany such decisions.
- Basic assumption: the decision problem is posed in probabilistic terms and all of the relevant probability values are known.
- **State of nature** = a variable ω which must be described probabilistically (e.g., in the fish example, $\omega = \omega_1$ for sea bass and $\omega = \omega_2$ for salmon).
- **Prior (probability)** = the prior knowledge about how likely the experimental result will be one or another before we can actually conduct the experiment (e.g., $P(\omega_1)$ and $P(\omega_2)$ depend upon the time of the year or the fishing area).
- **Decision rule** = Decide ω_1 if $P(\omega_1) > P(\omega_2)$, otherwise decide ω_2 (most basic case, when the decision should be made only upon the prior probabilities and under the assumption that any incorrect classification entails the same cost or consequence).

- **Conditional probability density function** $=$ the density function of a random variable whose distribution depends on the state of nature: $p(x|\omega_1), p(x|\omega_2)$, where x is an additional measurement meant to improve a classifier (i.e., the lightness of a fish).
- *How does some additional measurement and the prior probabilities influence our decision on a specific category?*

$$posterior = \frac{likelihood \times prior}{evidence} \qquad (Bayes' \ formula)$$

- The formula comes from the (joint) probability density of finding a pattern that is in category ω_j and has feature value x: $p(\omega_j, x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j)$, hence

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

- The formula shows how the prior probability $P(\omega_j)$, before anything is observed, is converted to a posterior probability $P(\omega_j|x)$ once observing the value of x.

- In general:
  1. the *likelihoods* (the category for which $p(x|\omega_j)$ is large is more "likely" to be correct) and the *prior* probabilities $P(\omega_j)$ are important in making a decision;
  2. the *evidence* is just a scale factor that states how frequently we will actually measure a pattern with feature value x.
- **Bayes' decision rule:** Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$, otherwise decide ω_2.
- The probability of error when making a decision: $P(error|x) = min\{P(\omega_1|x), P(\omega_2|x)\}$.

Continuous Features

- Generalizations:
  1. allowing the use of more than one feature
  2. allowing more than two states of nature
  3. allowing actions other than merely deciding the state of nature
  4. introducing a loss function more general than the probability of error
- $\omega_1, \ldots, \omega_c$ - the set of c states of nature
- $\alpha_1, \ldots, \alpha_a$ - the set of a possible actions
- $\lambda(\alpha_i|\omega_j) := \lambda_{ij}$ - the loss function
- $\mathrm{x} \in \mathbb{R}^d$ - the d-component feature vector
- $p(\mathrm{x}|\omega_j)$ - the probability density function for x conditioned on ω_j being the true state of nature.
- The posterior probability:

$$P(\omega_j|\mathrm{x}) = \frac{p(\mathrm{x}|\omega_j)P(\omega_j)}{p(\mathrm{x})}, \quad \text{where} \quad p(\mathrm{x}) = \sum_{j=1}^{c} p(\mathrm{x}|\omega_j)P(\omega_j).$$

- **Conditional risk** associated with taking action α_i:

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

- With a particular observation x, minimizing the expected loss implies selecting the action that minimizes the conditional risk.
- **Decision rule** $=$ a function $\alpha(\mathrm{x})$ that specifies which rule action to take for every possible observation.
- The overall risk:

$$R = \int R(\alpha(\mathrm{x})|\mathrm{x}) p(\mathrm{x}) d\mathrm{x}$$

- *Bayes decision rule* (reformulated): To minimize the overall risk, compute the conditional risk $R(\alpha_i|\mathrm{x})$ for $i = 1, \ldots, a$ and select the action for which the risk is minimum.

- The conditional risk is

$$R(\alpha_1|\mathrm{x}) = \lambda_{11}P(\omega_1|\mathrm{x}) + \lambda_{12}P(\omega_2|\mathrm{x})$$
$$R(\alpha_2|\mathrm{x}) = \lambda_{21}P(\omega_1|\mathrm{x}) + \lambda_{22}P(\omega_2|\mathrm{x})$$

- The minimum-risk decision rule in terms of posterior probabilities: decide ω_1 if

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

  in practice, the decision is generally determined by the more likely state of nature, although the posterior probabilities must be scaled by the loss differences.
- The decision rule in an equivalent form using the prior probabilities is: decide ω_1 if

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

- Another interpretation: decide ω_1 if

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

  that is, if the *likelihood ratio* $\frac{p(\mathrm{x}|\omega_1)}{p(\mathrm{x}|\omega_1)}$ is greater than a threshold value that is independent of the observation x.

- Usually, there is a connection between each state of nature and one of the actions, that is, if action α_i is taken and the true state of nature is ω_j, then the decision is correct if $i = j$ and in error if $i \neq j$.
- The *zero-one* loss function is applicable in this case (all errors are equally costly):

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \ldots, c.$$

- The risk is the average probability of error:

$$R(\alpha_i|\mathrm{x}) = \sum_{j \neq i} P(\omega_j|\mathrm{x}) = 1 - P(\omega_i|\mathrm{x}).$$

- Minimizing the risk means *maximizing* the posterior probability $P(\omega_i|\mathrm{x})$: decide ω_i if $P(\omega_i|\mathrm{x}) > P(\omega_j|\mathrm{x})$ for all $j \neq i$.

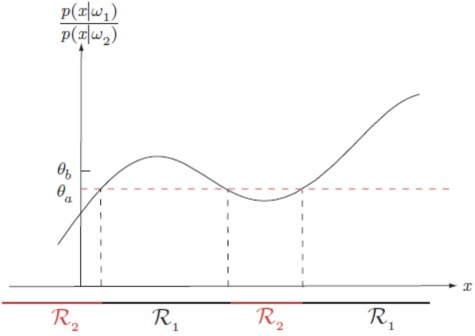

Figure: Likelihood ratio of two distributions. Here, θ_a corresponds to the zero-one loss, while θ_b corresponds to the situation when the loss function penalizes miscategorizing ω_2 as ω_1 more than the converse, that is, $\lambda_{12} > \lambda_{21}$.

Minimax Criterion

- There are situations when a classifier should perform well over a *range* of prior probabilities.
- The classifier should be designed to minimize the maximum possible overall risk (for any value of the priors).
- $\mathcal{R}_i$ - the region of the feature space where the classifier decides ω_i, $i = 1, 2$.
- The overall risk:

$$R = \int_{\mathcal{R}_1} [\lambda_{11} P(\omega_1) p(\mathrm{x}|\omega_1) + \lambda_{12} P(\omega_2) p(\mathrm{x}|\omega_2)] d\mathrm{x}$$
$$+ \int_{\mathcal{R}_2} [\lambda_{21} P(\omega_1) p(\mathrm{x}|\omega_1) + \lambda_{22} P(\omega_2) p(\mathrm{x}|\omega_2)] d\mathrm{x},$$

or, in terms of $P(\omega_1)$,

$$R(P(\omega_1)) = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathrm{x}|\omega_2) d\mathrm{x}$$

$$+P(\omega_1) \left[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathrm{x}|\omega_1) d\mathrm{x} - (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathrm{x}|\omega_2) d\mathrm{x} \right],$$

which shows that the overall risk is linear in $P(\omega_1)$ for determined $\mathcal{R}_1$ and $\mathcal{R}_2$.

$$
\begin{aligned}
R_{mm} &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathrm{x}|\omega_2) d\mathrm{x} \\
&= \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathrm{x}|\omega_1) d\mathrm{x}
\end{aligned}
$$

is the *minimax risk*, equal to the worst Bayes risk.

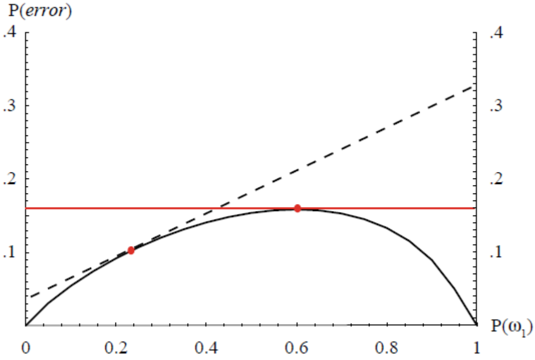

Figure: For a fixed optimal decision boundary, the probability of error will change as a linear function of $P(\omega_1)$. To minimize the maximum of such error, the decision boundary should be designed for the maximum Bayes error and thus the error will not change as a function of prior.

Classifiers, Discriminant Functions and Decision Surfaces

- An useful way to represent a pattern classifier is in terms of a set of *discriminant functions* $g_i(\mathrm{x})$, $i = 1, \ldots, c$.
- Such a classifier assigns a feature vector x to a class ω_i if $g_i(\mathrm{x}) > g_j(\mathrm{x})$ for all $j \neq i$.
- It can be viewed as a network or machine that computes c discriminant functions and selects the category corresponding to the largest discriminant.
- For the general case with risks, $g_i(\mathrm{x}) = R(\alpha_i|x)$.
- For the minimum-error-rate case, $g_i(\mathrm{x}) = P(\omega_i|\mathrm{x})$.
- For computational simplifications, note that a discriminant vector can be composed to any monotonically increasing function, without affecting the resulting classification.
- The effect of any decision rule is to divide the feature space into c *decision regions* $\mathcal{R}_1, \ldots, \mathcal{R}_c$. If $g_i(\mathrm{x}) > g_j(\mathrm{x})$ for any $j \neq i$, then x is in $\mathcal{R}_i$, therefore it should be assigned to ω_i.

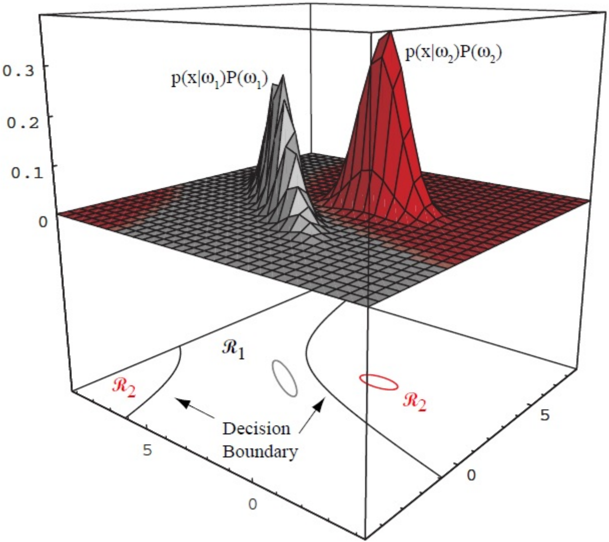

Figure: Two-dimensional two-category classifier with Gaussian probability densities.

The Normal (Gaussian) Univariate Density

- Density function:

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

- Expected value:

$$\mu = \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x)dx$$

- Variance (expected square deviation):

$$\sigma^2 = \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx$$

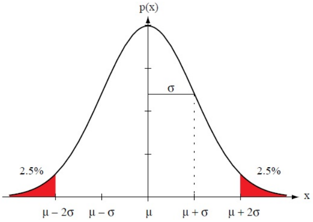

Figure: Roughly 95% of the area is in the range $|x-\mu| \leq 2\sigma$. The peak has value $p(\mu) = \frac{1}{\sqrt{2\pi}\sigma}$.

Multivariate Normal Density

- d-dimensional normal density:

$$p(\mathrm{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathrm{x}-\mu)^t \Sigma^{-1} (\mathrm{x}-\mu)\right] \sim N(\mu, \Sigma)$$

$$\mu = \mathcal{E}[\mathrm{x}] = \int \mathrm{x} p(\mathrm{x})$$

$$\Sigma = \mathcal{E}[(\mathrm{x}-\mu)(\mathrm{x}-\mu)^t] = \int (\mathrm{x}-\mu)(\mathrm{x}-\mu)^t p(\mathrm{x}) d\mathrm{x}$$

- Σ - symmetric and positive semidefinite; take the case when $|\Sigma| > 0$ (eliminate the case when sample vectors are drawn from a linear subspace).
- $\sigma_{ij} = 0 (\neq 0) \Rightarrow x_i, x_j$ are *statistically independent (correlated)*
- For a $d \times k$ matrix A and a k-vector $\mathrm{y} = A^y \mathrm{x}$, $p(\mathrm{y}) \sim N(A^t \mu, A^t \Sigma A)$.
- Knowledge of the covariance matrix allows the computation of the dispersion of the data in any direction, or in any subspace.
- *Spherical distribution* = a distribution having the covariance matrix proportional to the identity matrix I.

- *Whitening transformation:* a transformation A_w which makes the spectrum of eigenvectors of the transformed distribution uniform; e.g., $A_w = \Phi\Lambda^{\frac{1}{2}}\Phi^t$, where Φ is the matrix with columns the orthonormal eigenvectors of Σ and Λ is the diagonal matrix of eigenvalues.
- The transformed distribution has covariance matrix I (it is a circularly symmetric Gaussian).

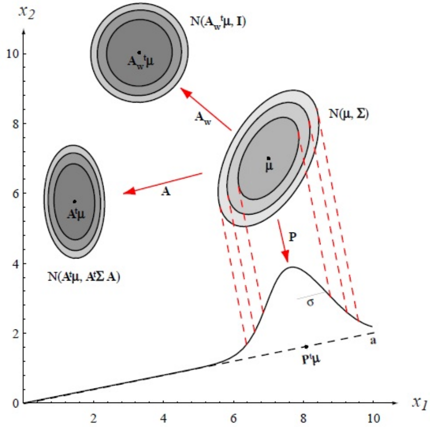

Figure: The action of a linear transformation on the feature space converts an arbitrary normal distribution into another normal distribution.