Pattern Classification

EET3053

Lecture 03: Bayesian Decision Theory

Dr. Kundan Kumar
Associate Professor
Department of ECE

Faculty of Engineering (ITER)
S'O'A Deemed to be University, Bhubaneswar, India-751030

Bayesian Decision Theory

- *Bayesian Decision Theory* is a fundamental statistical approach that quantifies the trade-offs between various decisions using probabilities and costs that accompany such decisions.
- First, we will assume that all probabilities are known.
- Then, we will study the cases where the probabilistic structure is not completely known.

Fish Sorting Example Revisited

- *State of nature* (class) is a random variable.
- Define ω as the type of fish we observe (state of nature, class) where
  - $\omega = \omega_1$ for sea bass,
  - $\omega = \omega_2$ for salmon.
  - $P(\omega_1)$ is the *a priori probability* that the next fish is a sea bass.
  - $P(\omega_2)$ is the *a priori probability* that the next fish is a salman.

Prior Probabilities

- Prior probabilities reflect our knowledge of how likely each type of fish will appear before we actually see it.
- How can we choose $P(\omega_1)$ and $P(\omega_2)$?
  - Set $P(\omega_1) = P(\omega_2)$ if they are equiprobable (uniform priors).
  - May use different values depending on the fishing area, time of the year, etc.
- Assume there are no other types of fish

$$\boxed{P(\omega_1) + P(\omega_2) = 1}$$

(exclusivity and exhaustivity)

Making a Decision

- How can we make a decision with only the prior information? (*Decision rule*)

$$\text{Decide} \begin{cases} \omega_1 & \text{if } P(\omega_1) > P(\omega_2) \\ \omega_2 & \text{otherwise} \end{cases}$$

- What is the *probability of error* for this decision?

$$P(error) = \min\{P(\omega_1), P(\omega_2)\}$$

- Don't you feel that there is some problem in making a decision?

Class-Conditional Probabilities

- Let's try to improve the decision using the lightness measurement x.
- Let x be a continuous random variable.
- Probability density function $p(x)$ (*evidence*)
  - how frequently we will measure a pattern with feature value x (e.g., x corresponds to lightness)
- Define $p(x|\omega_j)$ as the *class-conditional probability density*
  - how frequently we will measure a pattern with feature value x given that pattern belongs to class ω_j
- $p(x|\omega_1)$ and $p(x|\omega_2)$ describe the difference in lightness between populations of sea bass and salmon.

Class-Conditional Probabilities

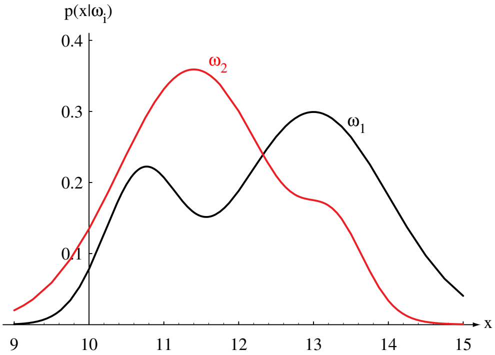


Figure: Hypothetical class-conditional probability density functions (lightness) for salmon/sea-bass

Posterior Probabilities

- Suppose we know $P(\omega_j)$ and $p(x|\omega_j)$ for $j = 1, 2$ and measure the lightness of a fish as the value x.
- Define $P(\omega_j|x)$ as the a *posterior probability* (probability of the state of nature being ω_j given the measurement of feature value x)
- We can use the *Bayes formula* to convert the prior probability to the posterior probability

$$\boxed{P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)} = \frac{likelihood \times prior}{evidence}}$$

where $p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j)$

Posterior Probabilities

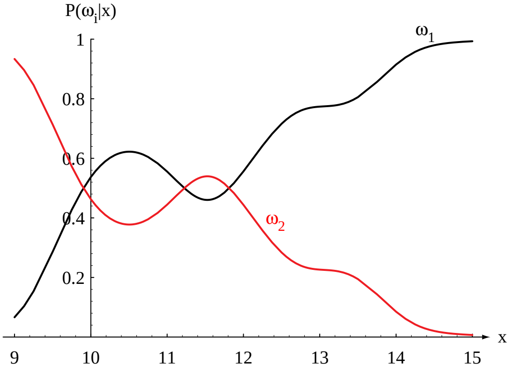


Figure: Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0.

Making a Decision

- $p(x|\omega_j)$ is called the *likelihood* and $p(x)$ is called the *evidence*.
- How can we make a decision after observing the value of x?

$$\text{Decide} \quad \begin{cases} \omega_1 & \text{if } P(\omega_1|x) > P(\omega_2|x) \\ \omega_2 & \text{otherwise} \end{cases}$$

- Rewriting the rule gives

$$\text{Decide} \quad \begin{cases} \omega_1 & \text{if } p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2) \\ \omega_2 & \text{otherwise} \end{cases}$$

- Note that, at every x, $P(\omega_1|x) + P(\omega_2|x) = 1$

Probability of Error

- What is the probability of error for this decision?

$$P(error|x) = \begin{cases} P(\omega_1|x) & \text{if we decide } \omega_2 \\ P(\omega_2|x) & \text{if we decide } \omega_1 \end{cases}$$

- What is the average probability of error?

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error|x)p(x)dx$$

- Bayes decision rule minimizes this error because

$$P(error|x) = \min\{P(\omega_1|x), P(\omega_2|x)\}$$

Generalization of the preceding ideas

Generalization of Bayes decision rule

- Use of more than one feature, e.g., $\{x_1,\ x_2, \ldots, x_d\}$
- Use more than two states of nature, e.g., $\{\omega_1,\ \omega_2, \ldots, \omega_c\}$
- Allowing actions and not only decide on the state of nature
  - take an action from the set of predefined actions $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$.
- Introduce a loss of function which is more general than the probability of error
  - Loss incurred $\lambda(\alpha_i|\omega_j)$ for taking action α_i while the true state of nature is ω_j.

Generalization of the preceding ideas

- Allowing the use of more than one feature merely requires replacing the scalar x by the feature vector x, where x is in a *d-dimensional Euclidean space*, $\mathbb{R}^d$, called the *feature space*.
- Allowing actions other than classification primarily allows the *possibility of rejection* – that is, of refusing to make a decision in close cases.
- The *loss function* states exactly how costly each action is, and is used to convert a probability determination into a decision.

Bayesian Decision Theory – Continuous Features

- Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the finite set of c states of nature (or "*classes*", "*categories*")
- Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the finite set of 'a' possible *actions*.
- Let $\lambda(\alpha_i|\omega_j)$ be the *loss* incurred for taking action α_i when the state of nature is ω_j.
- Let x be the d-component vector-valued random variable called the *feature vector*.

Bayesian Decision Theory – Continuous Features

- $p(\mathrm{x}|\omega_j)$ is the class-conditional probability density function.
- $P(\omega_j)$ is the prior probability that nature is in state ω_j.
- The posterior probability can be computed as

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

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

Conditional Risk

- Suppose we observe x and take action α_i.
- If the true state of nature is ω_j, we incur the loss $\lambda(\alpha_i|\omega_j)$.
- The expected loss with taking action α_i is

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

which is also called the *conditional risk*.

Minimum-Risk Classification

- The general *decision rule* $\alpha(\mathrm{x})$ tells us which action to take for observation x.
- We want to find the decision rule that minimizes the overall risk

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

- Bayes decision rule minimizes the overall risk by selecting the action α_i for which $R(\omega_i|\mathrm{x})$ is minimum.
- The resulting minimum overall risk is called the *Bayes risk* and is the best performance that can be achieved.

Two-Category Classification

- α_1 deciding true state of nature is ω_1.
  α_2 deciding true state of nature is ω_2.
  $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ = loss incurred for deciding ω_i when the true state of nature is ω_j.
- Conditional risk:

$$\begin{aligned} R(\alpha_1|\mathbf{x}) &= \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x}) \qquad \text{and} \\ R(\alpha_2|\mathbf{x}) &= \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x}). \end{aligned}$$

- Fundamental rule to decide ω_1, $R(\alpha_1|\mathrm{x}) < R(\alpha_2|\mathrm{x})$
- In terms of the posterior probabilities, decide ω_1 if

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

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

and decide ω_2 otherwise

Two-Category Classification

- the preceding rule is equivalent to the following rule:

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

  This is called *likelihood ratio*.
- Optimal decision property:
  "If the likelihood ratio exceeds a threshold value independent of the input pattern x, we can take optimal actions"

Minimum-Error-Rate Classification

- Classification: actions are decision on classes
  - If action α_i is taken and the true state of nature is ω_j then then decision is correct if $i = j$ and in error if $i \neq j$
- Seek a decision rule that minimizes the *probability of error* which is the *error rate*.

Minimum-Error-Rate Classification

- Define the *zero-one loss function*

$$\lambda\left(\alpha_i|\omega_j\right) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases} \quad i, j = 1, \ldots, c$$

- Conditional risk becomes

$$\begin{aligned} R\left(\alpha_i|\mathrm{x}\right) &= \sum_{j=1}^{c} \lambda\left(\alpha_i|\omega_j\right) P\left(\omega_j|\mathrm{x}\right) \\ &= \sum_{j \neq i} P(\omega_j|\mathrm{x}) \\ &= 1 - P(\omega_i|\mathrm{x}) \end{aligned}$$

Minimum-Error-Rate Classification

- Minimizing the risk requires maximizing $P(\omega_i|\mathrm{x})$ and results in the minimum-error decision rule

$$\boxed{\text{Decide } \omega_i \text{ if } P(\omega_i|\mathrm{x}) > P(\omega_j|\mathrm{x}) \quad \forall\, j \neq i.}$$

- The resulting error is called the *Bayes error* and is the best performance that can be achieved.

Minimum-Error-Rate Classification

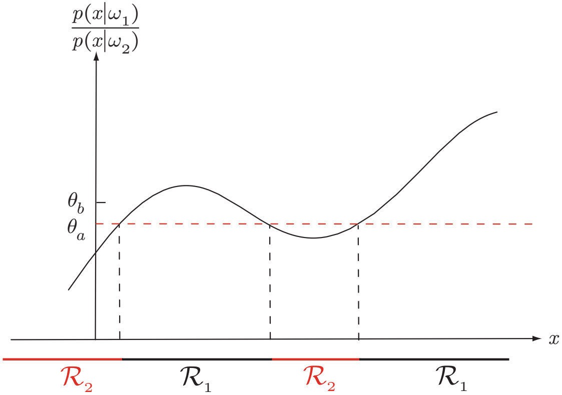


Figure: The likelihood ratio $p(\mathrm{x}|\omega_1)/p(\mathrm{x}|\omega_2)$. The threshold θ_a is computed using the priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$, and a zero-one loss function. If we penalize mistakes in classifying ω_2 patterns as ω_1 more than the converse, we should increase the threshold to θ_b.

Classifiers, Discriminant Functions, and Decision Surfaces

Classifiers

- There are many different ways to represent patterns classifiers.

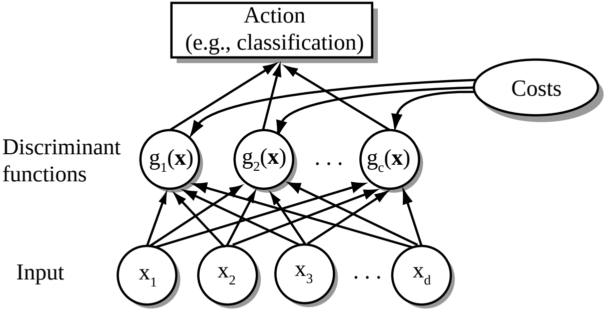


Figure: The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathrm{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly.

Discriminant Functions

- A useful way of representing classifiers is through *discriminant functions* $g_i(\mathrm{x}), i = 1, \ldots, c$, where the classifier assigns a feature vector x to class ω_i if

$$\boxed{g_i(\mathrm{x}) > g_j(\mathrm{x}) \quad \forall\ j \neq i.}$$

- For the classifier that minimizes conditional risk

$$\boxed{g_i(\mathrm{x}) = -R(\alpha_i|\mathrm{x}).}$$

- For the classifier that minimizes error

$$\boxed{g_i(\mathrm{x}) = P(\omega_i|\mathrm{x}).}$$

Discriminant Functions

- These functions divide the feature space into c *decision regions* $(\mathcal{R}_1, \ \ldots, \mathcal{R}_c)$, separated by *decision boundaries*.
- Note that the results do not change even if we replace every $g_i(\mathrm{x})$ by $f(g_i(\mathrm{x}))$ where $f(\cdot)$ is a monotonically increasing function (e.g., logarithm).
- This may lead to significant analytical and computational simplifications.

For example: Minimum-Error-Rate Classification

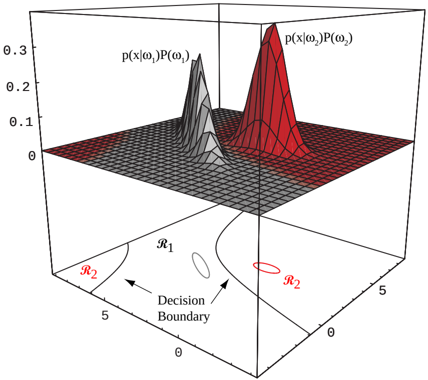


Figure: In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas.

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)}$$

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

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

Decision boundary: Two-Category Case

- The two-category case is just a special instance of the multicategory case.
- Instead of using two discriminant functions g_1 and g_2 and assigning x to ω_1 if $g_1 > g_2$, it is common to define a single discriminant function

$$\boxed{g(\mathrm{x}) \equiv g_1(\mathrm{x}) - g_2(\mathrm{x})}$$

and Decide ω_1 if $g(\mathrm{x}) > 0$; otherwise decide ω_2

- Minimum-error-rate discriminant function can be written as

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x})$$

$$g(\mathbf{x}) = \ln\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln\frac{P(\omega_1)}{P(\omega_2)}.$$

Normal/Gaussian Density

The Normal/Gaussian Density

- Univariate density, $N(\mu, \sigma^2)$
  - Density which is analytically tractable
  - Continuous density
  - A lot of processes are asymptotically Gaussian
  - Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)
  - For $x \in \mathbb{R}$:

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

where
$$\begin{aligned} \mu &= \text{mean (or expected value) of } x \\ &= E[x] = \int xp(x)dx \\ \sigma^2 &= \text{expected squared deviation or variance} \\ &= E[(x-\mu)(x-\mu)^t] = \int (x-\mu)(x-\mu)^t p(x)dx \end{aligned}$$

Univariate density

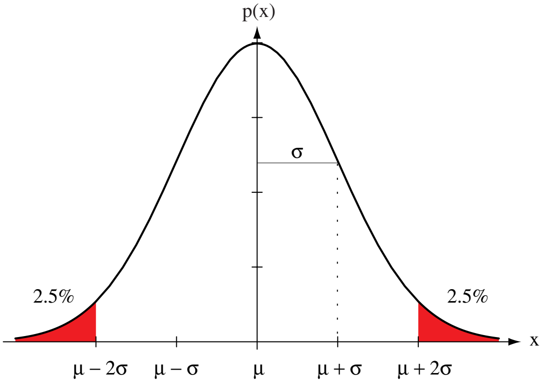


Figure: A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$

Multivariate Density

- Multivariate normal density, $N(\mu, \Sigma)$, in d-dimensions (i.e., for $\mathrm{x} \in \mathbb{R}^d$) is

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

where:

$\mathrm{x} = (x_1, x_2, \ldots, x_d)^T$ d-dimensional vector

$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^T$ mean vector

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

$\Sigma = d \times d$ covariance matrix

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

$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively

Multivariate Density

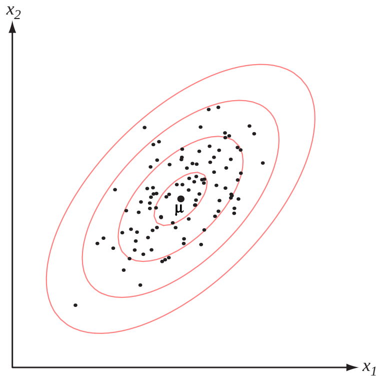


Figure: Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean μ. The loci of points of constant density are the ellipses for which $(\mathrm{x}-\mu)^t\Sigma^{-1}(\mathrm{x}-\mu)$ is constant, where the eigenvectors of Σ determine the direction and the corresponding eigenvalues determine the length of the principal axes. The quantity $r^2 = (\mathrm{x}-\mu)^t\Sigma^{-1}(\mathrm{x}-\mu)$ is called the squared *Mahalanobis distance* from x to μ

Discriminant Functions for the Normal Density

- Discriminant functions for minimum-error-rate classification can be written as

$$\boxed{g_i(\mathrm{x}) = \ln \mathrm{p}(\mathrm{x}|\omega_\mathrm{i}) + \ln \mathrm{P}(\omega_\mathrm{i})}$$

- For $p(\mathrm{x}|\omega_\mathrm{i}) = N(\mu_i, \Sigma_i)$ (case of multivariate normal)

$$\boxed{g_i(\mathrm{x}) = -\frac{1}{2}(\mathrm{x} - \mu_i)^T \Sigma_i^{-1} (\mathrm{x} - \mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)}$$

Case 1: $\Sigma_i = \sigma^2 I$

- Discriminant functions are

$$g_i(\mathrm{x}) = \mathrm{w}_i^T \mathrm{x} + w_{i0} \qquad \text{linear discriminant}$$

where

$$\mathrm{w}_i = \frac{1}{\sigma^2}\mu_i$$

$$w_{i0} = -\frac{1}{2\sigma^2}\mu_i^T \mu_i + \ln P(w_i)$$

(w_{i0} is the threshold or bias for the ith category)

Case 1: $\Sigma_i = \sigma^2 I$

- Decision boundaries are the hyperplanes $g_i(\mathrm{x}) = g_j(\mathrm{x})$, and can be written as

$$\mathrm{w}^T(\mathrm{x} - \mathrm{x}_0)$$

where

$$\begin{aligned} \mathrm{w} &= \mu_i - \mu_j \\ \mathrm{x}_0 &= \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{||\mu_i - \mu_j||^2} \ln \frac{P(w_i)}{P(w_j)}(\mu_i - \mu_j). \end{aligned}$$

- Hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ passes through the point x_0 and is orthogonal to the vector w.

Case 1: $\Sigma_i = \sigma^2 I$

- If the covariances of two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $(d-1)$ dimensions, perpendicular to the line separating the means.

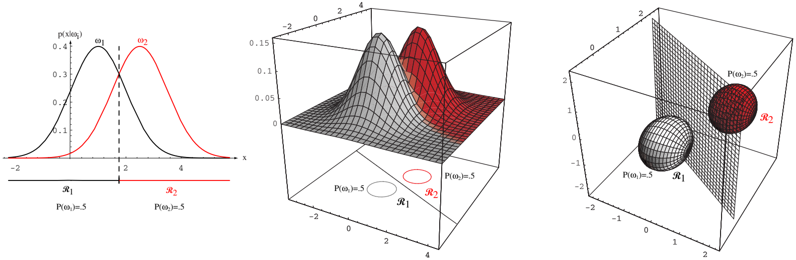
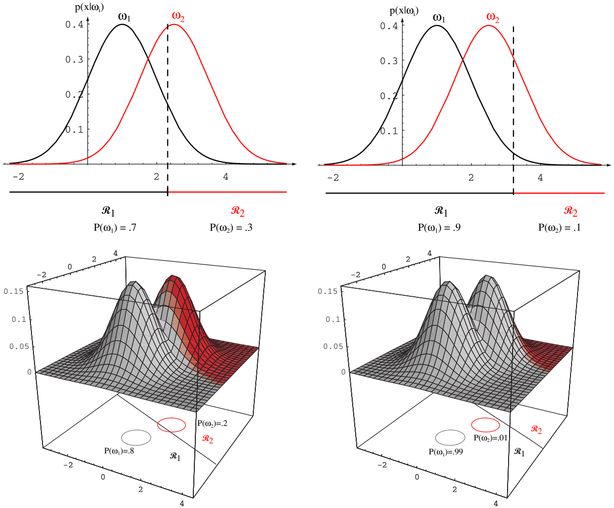


Figure: In these 1-, 2-, and 3-dimensional examples, we indicate $p(\mathrm{x}|w_i)$ and the boundaries for the case $P(w_1) = P(w_2)$. In this 3-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$.

Case 1: $\Sigma_i = \sigma^2 I$



Case 1: $\Sigma_i = \sigma^2 I$

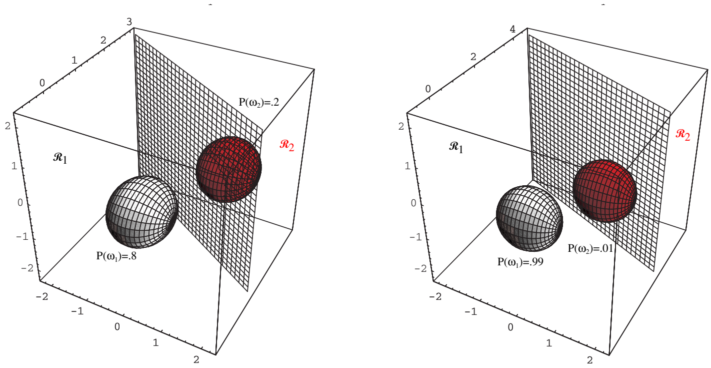


Figure: A the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these 1-, 2-, and 3-dimensional spherical Gaussian distributions.

Case 1: $\Sigma_i = \sigma^2 I$

- Special case when $P(w_i)$ are the same for $i = 1, \dots, c$ is the minimum-distance classifier that uses the decision rule

$$\text{assign } \mathrm{x} \text{ to } w_{i^*} \quad \text{where} \quad i^* = \arg\min_{i=1,\dots,c} ||\mathrm{x} - \mu_i||$$

Case 2: $\Sigma_i = \Sigma$

- Discriminant functions are

$$g_(\mathrm{x}) = \mathrm{w}_i^T \mathrm{x} + w_{i0} \qquad \text{(linear discriminant)}$$

where

$$\begin{aligned} \mathrm{w}_i =& \Sigma^{-1}\mu_i \\ w_{i0} =& -\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \ln P(w_i). \end{aligned}$$

Case 2: $\Sigma_i = \Sigma$

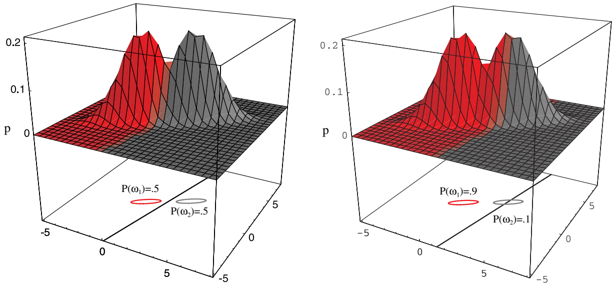
- Decision boundaries can be written as

$$\mathrm{w}^T(\mathrm{x} - \mathrm{x}_0) = 0$$

$$\begin{aligned} \mathrm{w} =& \Sigma^{-1}(\mu_i - \mu_j) \\ \mathrm{x}_0 =& \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln(P(w_i)/P(w_j))}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j). \end{aligned}$$

- Hyperplane passes through x_0 but is not necessarily orthogonal to the line between the means.

Case 2: $\Sigma_i = \Sigma$



Case 2: $\Sigma_i = \Sigma$

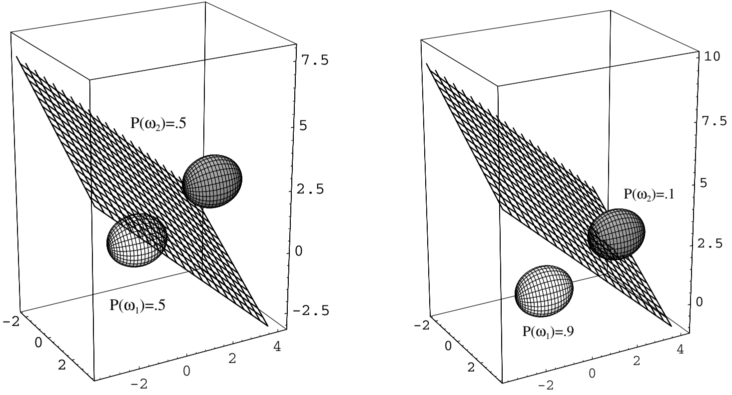


Figure: Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means.

Case 3: Σ_i =arbitrary

- Discriminant functions are

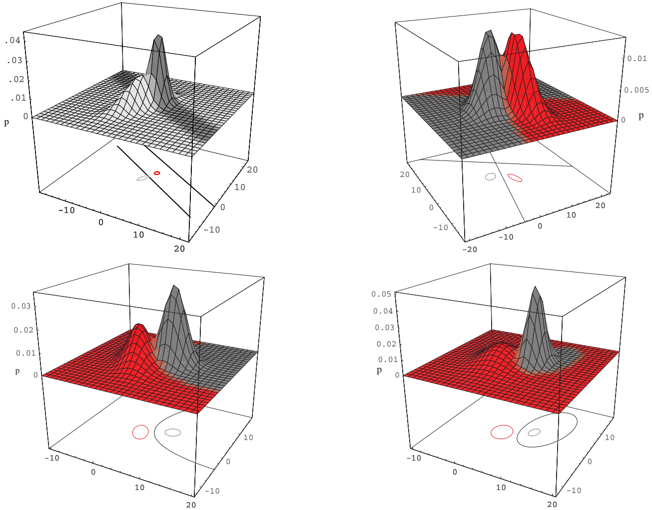
$$g_i(\mathrm{x}) = \mathrm{x}^T \mathrm{W}_i \mathrm{x} + \mathrm{w}_i^T \mathrm{x} + w_{i0} \qquad \text{(quadratic discriminant)}$$

where

$$\begin{aligned} \mathrm{W}_i &= -\frac{1}{2}\Sigma_i^{-1} \\ \mathrm{w}_i &= \Sigma_i^{-1}\mu_i \\ w_{i0} &= -\frac{1}{2}\mu_i^T \Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i) \end{aligned}$$

- Decision boundaries are hyperquadrics.

Case 3: Σ_i =arbitrary



Case 3: Σ_i =arbitrary

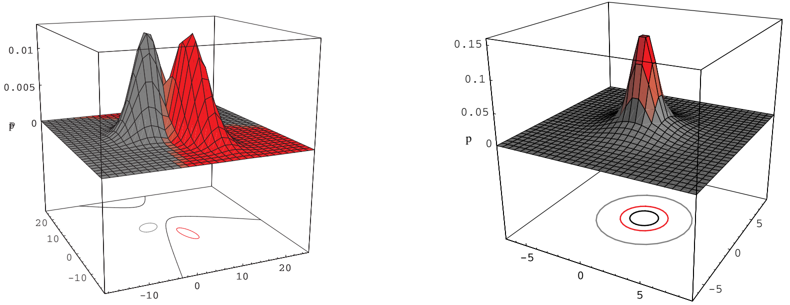


Figure: Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadratic, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric.

Example to solve

Question:

For a 2-class problem, the prior probabilities are: $P(w_1) = 1/4$ and $P(w_2) = 3/4$. The class conditional distribution for $\mathrm{x} = x$, that is x has only a single attribute, are $p(x/w_1) = N(0,1)$ and $p(x/w_2) = N(1,1)$.

(a) Calculate the threshold boundary value x_t which gives the probability of minimum error.

(b) If the loss matrix is

$$\lambda_{ij} = \begin{bmatrix} 0 & 1 \\ 1/2 & 0 \end{bmatrix},$$

find the threshold boundary value x_t for minimum risk.

Example to solve

Question:

Two normal distribution are characterized by: $P(w_1) = P(w_2) = 0.5$ and

$$\mu_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \mu_2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

Sketch the Bayes decision boundary for $\Sigma_1 = \Sigma_2 = I$.

Example to solve

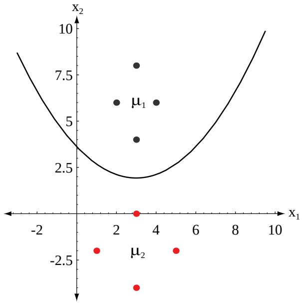
Question:

Find the decision boundary between ω_1 and ω_2 where

$$\omega_1 : \begin{pmatrix} 2 \\ 6 \end{pmatrix}, \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \begin{pmatrix} 3 \\ 8 \end{pmatrix}, \begin{pmatrix} 4 \\ 6 \end{pmatrix}$$

$$\omega_2 : \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \begin{pmatrix} 3 \\ -4 \end{pmatrix}, \begin{pmatrix} 5 \\ -2 \end{pmatrix}$$

$$\& \quad P(\omega_1) = P(\omega_2) = 0.5$$



Assuming that samples in ω_1 and ω_2 following Normal distribution. *Solution:*
$1.5x_1^2 - 9x_1 - 8x_2 + 28.1137 = 0$

Evaluate Classifiers

Confusion Matrix

- For a two class-problem, a table of confusion (sometimes also called a confusion matrix), is a table with two rows and two columns that reports the number of
  - *false positives (FP)*,
  - *false negatives (FN)*,
  - *true positives (TP)*, and
  - *true negatives(TN)*
- In statistical classification, a confusion matrix, also known as an error matrix.

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Performance Evaluation using confusion matrix

- True positive rate (TPR), also called Sensitivity
- False positive rate (FPR), also called Fall-out
- False negative rate (FNR), also called Miss rate
- True negative rate (TNR), also called Specificity

True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$
False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$

Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$

Receiver Operating Characteristics

- If we use a parameter (e.g., a threshold) in our decision, the plot of TPR vs FPR for different values of the parameter is called the *receiver operating characteristic (ROC)* curve.
- The ROC curve is created by plotting the *true positive rate (TPR)* against the *false positive rate (FPR)* at various threshold settings.

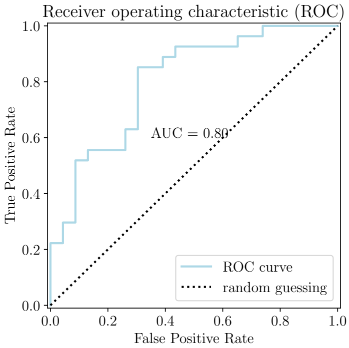
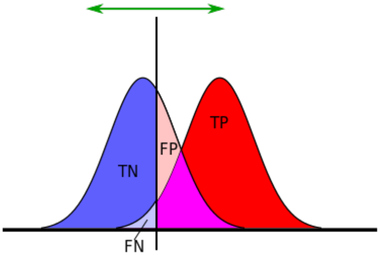
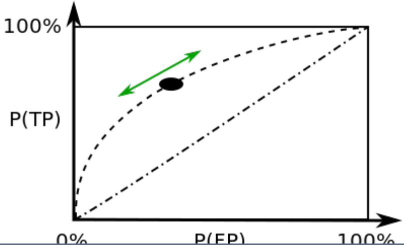


Figure: Example receiver operating characteristic (ROC) curves for different setting of the system

Receiver Operating Characteristics



TP	FP
FN	TN



Summary

- To minimize the overall risk, choose the action that minimizes the conditional risk $R(\alpha|\mathrm{x})$.
- To minimize the probability of error, choose the class that maximizes the posterior probability $P(\omega_j|\mathrm{x})$.
- If there are different penalties for misclassifying patterns from different classes, the posteriors must be weighted according to such penalties before taking action.
- Do not forget that these decisions are the optimal ones under the assumption that the "true" values of the probabilities are known.

Bayes Decision Theory - Discrete Features

Bayes Decision Theory - Discrete Features

- Components of x are binary or integer valued, x can take only one of m discrete values $\mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_m$
- Case of independent binary features in 2 category problem
  Let $\mathrm{x} = [x_1, x_2, \ldots, x_d]^t$ where each x_i is either 0 or 1, with probabilities:
  $$p_i = P(x_i = 1|\omega_1)$$
  $$q_i = P(x_i = 1|\omega_2)$$
  $p_i > q_i \Rightarrow x_i$ is more likely to have value 1 if $\mathrm{x} \in \omega_1$
- Class conditional probabilities

$$\boxed{p(\mathrm{x}|\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i}}$$

$$\boxed{p(\mathrm{x}|\omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1-x_i}}$$

Bayes Decision Theory - Discrete Features

- Then the likelihood ratio is given by

$$\frac{p(\mathrm{x}|\omega_1)}{p(\mathrm{x}|\omega_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

- we know that

$$g(\mathbf{x}) = \ln\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln\frac{P(\omega_1)}{P(\omega_2)}$$

- Therefore discriminant function will be

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

Bayes Decision Theory - Discrete Features

- We note especially that this discrimant function is linear in the x_i and thus we can write

$$g(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + w_0,$$

where

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1, ..., d$$

and

$$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}.$$

- Decide ω_1 if $g(\mathrm{x}) > 0$ and ω_2 if $g(\mathrm{x}) \leq 0$

Bayes Decision Theory - Discrete Features

- If $p_i = q_i$, x_i gives us no information about the state of nature, and ω_0.
- If $p_i > q_i$, then $1 - p_i < 1 - q_i$ and w_i is positive. Thus in this case a "yes" answer for x_i contribute w_i votes for ω_1.
- Furthermore, for any fixed $q_i < 1$, w_i gets larger as p_i gets larger.
- On the other hand, if $p_i < q_i$, w_i is negative and a "yes" answer contributes $|w_i|$ votes for ω_2.

Example to solve

Question:

Compute Bayesian decision for three-dimensional binary features

Suppose two categories consist of independent binary features in three dimensions with known feature probabilities. Let us construct the Bayesian decision boundary if $P(\omega_1) = P(\omega_2) = 0.5$ and the individual components obey:

$$\begin{cases} p_i = 0.8 \\ q_i = 0.5 \end{cases} \qquad i = 1, 2, 3$$

Example to solve

Question:

Compute Bayesian decision for three-dimensional binary features

Suppose two categories consist of independent binary features in three dimensions with known feature probabilities. Let us construct the Bayesian decision boundary if $P(\omega_1) = P(\omega_2) = 0.5$ and the individual components obey:

$$\begin{cases} p_1 = p_2 = 0.8, \ p_3 = 0.5 \\ q_1 = q_2 = q_3 = 0.5 \end{cases}$$

Addition Examples:

Question:

13. In many pattern classification problems one has the option either to assign the pattern to one of c classes, or to *reject* it as being unrecognizable. If the cost for rejects is not too high, rejection may be a desirable action. Let

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \quad i,j = 1,...,c \\ \lambda_r & i = c+1 \\ \lambda_s & \text{otherwise,} \end{cases}$$

where λ_r is the loss incurred for choosing the $(c+1)$th action, rejection, and λ_s is the loss incurred for making a substitution error. Show that the minimum risk is obtained if we decide ω_i if $P(\omega_i|\mathbf{x}) \geq P(\omega_j|\mathbf{x})$ for all j and if $P(\omega_i|\mathbf{x}) \geq 1 - \lambda_r/\lambda_s$, and reject otherwise. What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$?

Question:

What is the inverse of $\begin{bmatrix} 4 & 4 \\ 1 & 2 \end{bmatrix}$?

References

[1] Hart, P. E., Stork, D. G., & Duda, R. O. (2000). Pattern classification. Hoboken: Wiley.

[2] Gose, E. (1997). Pattern recognition and image analysis.

Thank you!