Bayesian Decision Theory

Selim Aksoy

Department of Computer Engineering
Bilkent University
saksoy@cs.bilkent.edu.tr

CS 551, Fall 2018
Bayesian Decision Theory is a fundamental statistical approach that quantifies the tradeoffs 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.
State of nature is a random variable.

Define $w$ as the type of fish we observe (state of nature, class) where

- $w = w_1$ for sea bass,
- $w = w_2$ for salmon.

$P(w_1)$ is the a priori probability that the next fish is a sea bass.

$P(w_2)$ is the a priori probability that the next fish is a salmon.
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 $P(w_1)$ and $P(w_2)$?
  - Set $P(w_1) = P(w_2)$ 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

$$P(w_1) + P(w_2) = 1$$

(exclusivity and exhaustivity).
Making a Decision

- How can we make a decision with only the prior information?

\[
\text{Decide } \begin{cases} 
  w_1 & \text{if } P(w_1) > P(w_2) \\
  w_2 & \text{otherwise}
\end{cases}
\]

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

\[
P(\text{error}) = \min\{P(w_1), P(w_2)\}
\]
Let’s try to improve the decision using the lightness measurement $x$.

Let $x$ be a continuous random variable.

Define $p(x|w_j)$ as the **class-conditional probability density** (probability of $x$ given that the state of nature is $w_j$ for $j = 1, 2$).

$p(x|w_1)$ and $p(x|w_2)$ describe the difference in lightness between populations of sea bass and salmon.
Class-Conditional Probabilities

Figure 1: Hypothetical class-conditional probability density functions for two classes.
Posterior Probabilities

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

\[
P(w_j|x) = \frac{p(x|w_j)P(w_j)}{p(x)}
\]

where \( p(x) = \sum_{j=1}^{2} p(x|w_j)P(w_j) \).
Making a Decision

- $p(x|w_j)$ is called the **likelihood** and $p(x)$ is called the **evidence**.

- How can we make a decision after observing the value of $x$?

  Decide \[
  \begin{aligned}
  w_1 & \text{ if } P(w_1|x) > P(w_2|x) \\
  w_2 & \text{ otherwise}
  \end{aligned}
  \]

- Rewriting the rule gives

  Decide \[
  \begin{aligned}
  w_1 & \text{ if } \frac{p(x|w_1)}{p(x|w_2)} > \frac{P(w_2)}{P(w_1)} \\
  w_2 & \text{ otherwise}
  \end{aligned}
  \]

- Note that, at every $x$, $P(w_1|x) + P(w_2|x) = 1$. 
Probability of Error

What is the probability of error for this decision?

\[ P(error|x) = \begin{cases} P(w_1|x) & \text{if we decide } w_2 \\ P(w_2|x) & \text{if we decide } w_1 \end{cases} \]

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

\[ P(error|x) = \min\{P(w_1|x), P(w_2|x)\}. \]
Bayesian Decision Theory

- How can we generalize to
  - more than one feature?
    - replace the scalar $x$ by the feature vector $\mathbf{x}$
  - more than two states of nature?
    - just a difference in notation
  - allowing actions other than just decisions?
    - allow the possibility of rejection
  - different risks in the decision?
    - define how costly each action is
Let \( \{w_1, \ldots, w_c\} \) be the finite set of \( c \) states of nature (classes, categories).

Let \( \{\alpha_1, \ldots, \alpha_a\} \) be the finite set of \( a \) possible actions.

Let \( \lambda(\alpha_i | w_j) \) be the loss incurred for taking action \( \alpha_i \) when the state of nature is \( w_j \).

Let \( x \) be the \( d \)-component vector-valued random variable called the feature vector.
Bayesian Decision Theory

- \( p(x|w_j) \) is the class-conditional probability density function.
- \( P(w_j) \) is the prior probability that nature is in state \( w_j \).
- The posterior probability can be computed as

\[
P(w_j|x) = \frac{p(x|w_j)P(w_j)}{p(x)}
\]

where \( p(x) = \sum_{j=1}^{c} p(x|w_j)P(w_j). \)
Conditional Risk

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

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

which is also called the conditional risk.
The general decision rule $\alpha(x)$ tells us which action to take for observation $x$.

We want to find the decision rule that minimizes the overall risk

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

Bayes decision rule minimizes the overall risk by selecting the action $\alpha_i$ for which $R(\alpha_i|x)$ is minimum.

The resulting minimum overall risk is called the Bayes risk and is the best performance that can be achieved.
Two-Category Classification

- Define
  - $\alpha_1$: deciding $w_1$,
  - $\alpha_2$: deciding $w_2$,
  - $\lambda_{ij} = \lambda(\alpha_i | w_j)$.

- Conditional risks can be written as

$$R(\alpha_1 | x) = \lambda_{11} P(w_1 | x) + \lambda_{12} P(w_2 | x),$$

$$R(\alpha_2 | x) = \lambda_{21} P(w_1 | x) + \lambda_{22} P(w_2 | x).$$
The *minimum-risk decision rule* becomes

\[
\begin{aligned}
\text{Decide} & \quad \begin{cases}
  w_1 & \text{if } (\lambda_{21} - \lambda_{11})P(w_1|x) > (\lambda_{12} - \lambda_{22})P(w_2|x) \\
  w_2 & \text{otherwise}
\end{cases}
\end{aligned}
\]

This corresponds to deciding \( w_1 \) if

\[
\frac{p(x|w_1)}{p(x|w_2)} > \frac{(\lambda_{12} - \lambda_{22}) P(w_2)}{(\lambda_{21} - \lambda_{11}) P(w_1)}
\]

\( \Rightarrow \) comparing the *likelihood ratio* to a threshold that is independent of the observation \( x \).
Minimum-Error-Rate Classification

- Actions are decisions on classes ($\alpha_i$ is deciding $w_i$).
- If action $\alpha_i$ is taken and the true state of nature is $w_j$, then the decision is correct if $i = j$ and in error if $i \neq j$.
- We want to find a decision rule that minimizes the probability of error.
Define the **zero-one loss function**

\[
\lambda(\alpha_i | w_j) = \begin{cases} 
0 & \text{if } i = j \\
1 & \text{if } i \neq j
\end{cases}
\]

\[i, j = 1, \ldots, c\]

(all errors are equally costly).

Conditional risk becomes

\[
R(\alpha_i | x) = \sum_{j=1}^{c} \lambda(\alpha_i | w_j) P(w_j | x)
\]

\[= \sum_{j \neq i} P(w_j | x)\]

\[= 1 - P(w_i | x).\]
Minimizing the risk requires maximizing $P(w_i|x)$ and results in the *minimum-error decision rule*

\[
\text{Decide } w_i \text{ if } P(w_i|x) > P(w_j|x) \quad \forall j \neq i.
\]

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

Figure 2: The likelihood ratio \( \frac{p(x|w_1)}{p(x|w_2)} \). The threshold \( \theta_a \) is computed using the priors \( P(w_1) = \frac{2}{3} \) and \( P(w_2) = \frac{1}{3} \), and a zero-one loss function. If we penalize mistakes in classifying \( w_2 \) patterns as \( w_1 \) more than the converse, we should increase the threshold to \( \theta_b \).
Discriminant Functions

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

\[
g_i(x) > g_j(x) \quad \forall j \neq i.
\]

- For the classifier that minimizes conditional risk

\[
g_i(x) = -R(\alpha_i | x).
\]

- For the classifier that minimizes error

\[
g_i(x) = P(w_i | x).
\]
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 $g_i(x)$ by $f(g_i(x))$ where $f(\cdot)$ is a monotonically increasing function (e.g., logarithm).

- This may lead to significant analytical and computational simplifications.
The Gaussian Density

- Gaussian can be considered as a model where the feature vectors for a given class are continuous-valued, randomly corrupted versions of a single typical or prototype vector.

- Some properties of the Gaussian:
  - Analytically tractable.
  - Completely specified by the 1st and 2nd moments.
  - Has the maximum entropy of all distributions with a given mean and variance.
  - Many processes are asymptotically Gaussian (Central Limit Theorem).
  - Linear transformations of a Gaussian are also Gaussian.
  - Uncorrelatedness implies independence.
Univariate Gaussian

▶ For $x \in \mathbb{R}$:

$$p(x) = N(\mu, \sigma^2)$$

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

where

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

$$\sigma^2 = E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 \ p(x) \ dx.$$
Figure 3: A univariate Gaussian distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$. 
Multivariate Gaussian

For $x \in \mathbb{R}^d$:

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

where

$$\mu = E[x] = \int x p(x) \, dx,$$

$$\Sigma = E[(x - \mu)(x - \mu)^T] = \int (x - \mu)(x - \mu)^T p(x) \, dx.$$
Figure 4: Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\mu$. The loci of points of constant density are the ellipses for which $(x - \mu)^T \Sigma^{-1} (x - \mu)$ is constant, where the eigenvectors of $\Sigma$ determine the direction and the corresponding eigenvalues determine the length of the principal axes. The quantity $r^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$ is called the squared *Mahalanobis distance* from $x$ to $\mu$. 

Multivariate Gaussian
Linear Transformations

- Recall that, given $x \in \mathbb{R}^d$, $A \in \mathbb{R}^{d \times k}$, $y = A^T x \in \mathbb{R}^k$, if $x \sim N(\mu, \Sigma)$, then $y \sim N(A^T \mu, A^T \Sigma A)$.

- As a special case, the whitening transform

$$A_w = \Phi \Lambda^{-1/2}$$

where

- $\Phi$ is the matrix whose columns are the orthonormal eigenvectors of $\Sigma$,

- $\Lambda$ is the diagonal matrix of the corresponding eigenvalues,

  gives a covariance matrix equal to the identity matrix $I$. 
Discriminant Functions for the Gaussian Density

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

\[ g_i(x) = \ln p(x|w_i) + \ln P(w_i). \]

For \( p(x|w_i) = N(\mu_i, \Sigma_i) \)

\[ g_i(x) = -\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(w_i). \]
Case 1: $\Sigma_i = \sigma^2 I$

- Discriminant functions are

$$g_i(x) = w_i^T x + w_{i0} \quad \text{(linear discriminant)}$$

where

$$w_i = \frac{1}{\sigma^2} \mu_i$$

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

($w_{i0}$ is the threshold or bias for the $i^{th}$ category).
Case 1: $\Sigma_i = \sigma^2 I$

- Decision boundaries are the hyperplanes $g_i(x) = g_j(x)$, and can be written as
  $$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

  where

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

- Hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ passes through the point $\mathbf{x}_0$ and is orthogonal to the vector $\mathbf{w}$. 
Case 1: $\Sigma_i = \sigma^2 I$

Figure 5: If the covariance matrices 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. The decision boundary shifts as the priors are changed.
Case 1: $\Sigma_i = \sigma^2 I$

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

$$\text{assign } x \text{ to } w_{i^*} \text{ where } i^* = \arg \min_{i=1,\ldots,c} \|x - \mu_i\|.$$
Case 2: $\Sigma_i = \Sigma$

- Discriminant functions are

$$g_i(x) = w_i^T x + w_{i0}$$  \hspace{1cm} \text{(linear discriminant)}

where

$$w_i = \Sigma^{-1} \mu_i$$

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

- Decision boundaries can be written as

$$w^T(x - x_0) = 0$$

where

$$w = \Sigma^{-1}(\mu_i - \mu_j)$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln(P(w_i)/P(w_j))}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j).$$

- Hyperplane passes through $x_0$ but is not necessarily orthogonal to the line between the means.
Case 2: $\Sigma_i = \Sigma$

Figure 6: Probability densities with equal but asymmetric Gaussian distributions. The decision hyperplanes are not necessarily perpendicular to the line connecting the means.
Case 3: $\Sigma_i = \text{arbitrary}$

- Discriminant functions are

$$g_i(x) = x^T W_i x + w_i^T x + w_{i0} \quad (\text{quadratic discriminant})$$

where

$$W_i = -\frac{1}{2} \Sigma_i^{-1}$$

$$w_i = \Sigma_i^{-1} \mu_i$$

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

- Decision boundaries are hyperquadrics.
Case 3: $\Sigma_i = \text{arbitrary}$

Figure 7: Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics.
Case 3: $\Sigma_i = \text{arbitrary}$

Figure 8: Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics.
For the two-category case

\[
P(\text{error}) = P(x \in \mathcal{R}_2, w_1) + P(x \in \mathcal{R}_1, w_2) \\
= P(x \in \mathcal{R}_2 | w_1)P(w_1) + P(x \in \mathcal{R}_1 | w_2)P(w_2) \\
= \int_{\mathcal{R}_2} p(x | w_1) P(w_1) \, dx + \int_{\mathcal{R}_1} p(x | w_2) P(w_2) \, dx.
\]
For the multicategory case

\[
P(error) = 1 - P(correct)
\]

\[
= 1 - \sum_{i=1}^{c} P(x \in R_i, w_i)
\]

\[
= 1 - \sum_{i=1}^{c} P(x \in R_i | w_i) P(w_i)
\]

\[
= 1 - \sum_{i=1}^{c} \int_{R_i} p(x | w_i) P(w_i) \, dx.
\]
Figure 9: Components of the probability of error for equal priors and the non-optimal decision point $x^*$. The optimal point $x_B$ minimizes the total shaded area and gives the Bayes error rate.
Consider the two-category case and define
- $w_1$: target is present,
- $w_2$: target is not present.

Table 1: *Confusion matrix*.

<table>
<thead>
<tr>
<th>Assigned</th>
<th>$w_1$</th>
<th>$w_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>correct detection</td>
<td>mis-detection</td>
</tr>
<tr>
<td>$w_1$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$w_2$</td>
<td>false alarm</td>
<td>correct rejection</td>
</tr>
</tbody>
</table>

- Mis-detection is also called false negative or Type II error.
- False alarm is also called false positive or Type I error.
If we use a parameter (e.g., a threshold) in our decision, the plot of these rates for different values of the parameter is called the **receiver operating characteristic (ROC) curve**.

**Figure 10**: Example receiver operating characteristic (ROC) curves for different settings of the system.
Summary

► To minimize the overall risk, choose the action that minimizes the conditional risk $R(\alpha|x)$.

► To minimize the probability of error, choose the class that maximizes the posterior probability $P(w_j|x)$.

► 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.