The Gaussian classifier

Nuno Vasconcelos

*ECE Department, UCSD*
Bayesian decision theory

recall that we have

- Y – state of the world
- X – observations
- g(x) – decision function
- L[g(x),y] – loss of predicting y with g(x)

Bayes decision rule is the rule that minimizes the risk

\[ Risk = E_{X,Y}[L(X,Y)] \]

given x, it consists of picking the prediction of minimum conditional risk

\[ g^*(x) = \arg \min_{g(x)} \sum_{i=1}^{M} P_{Y|X}(i \mid x) L[g(x), i] \]
MAP rule

▶ for the “0-1” loss

\[
L[g(x), y] = \begin{cases} 
1, & g(x) \neq y \\
0, & g(x) = y 
\end{cases}
\]

▶ the optimal decision rule is the maximum a-posteriori probability rule

\[
g^*(x) = \arg \max_{i} P_{Y|X}(i \mid x)
\]

▶ the associated risk is the probability of error of this rule (Bayes error)

▶ there is no other decision function with lower error
MAP rule

by application of simple mathematical laws (Bayes rule, monotonicity of the log)

we have shown that the following three decision rules are optimal and equivalent

• 1) \[ i^*(x) = \operatorname{arg\,max}_i P_{Y|X}(i \mid x) \]

• 2) \[ i^*(x) = \operatorname{arg\,max}_i \left[ P_{X|Y}(x \mid i) P_Y(i) \right] \]

• 3) \[ i^*(x) = \operatorname{arg\,max}_i \left[ \log P_{X|Y}(x \mid i) + \log P_Y(i) \right] \]

• 1) is usually hard to use, 3) is frequently easier than 2)
Example

- The Bayes decision rule is usually **highly intuitive**
- We have used an example from communications
  - A bit is transmitted by a source, corrupted by noise, and received by a decoder

![Diagram of a channel with inputs X and Y]

- Q: What should the **optimal decoder** do to recover Y?
Example

this was modeled as a classification problem with Gaussian classes

\[
P_{X|Y}(x \mid 0) = G(x, \mu_0, \sigma)
\]
\[
P_{X|Y}(x \mid 1) = G(x, \mu_1, \sigma)
\]

• or, graphically,
BDR

for which the optimal decision boundary is a threshold

- pick “0” if

\[
x < \frac{\mu_1 + \mu_0}{2} + \frac{\sigma^2}{\mu_1 - \mu_0} \log \frac{P_Y(0)}{P_Y(1)}
\]
back to our signal decoding problem

- in this case $T = 0.5$
- decision rule

\[
Y = \begin{cases} 
0, & \text{if } x < 0.5 \\
1, & \text{if } x > 0.5 
\end{cases}
\]

- this is intuitive
- we place the threshold midway along the noise sources and adapt according to the class prior
what is the role of the prior for class probabilities?

\[ x < \frac{\mu_1 + \mu_0}{2} + \frac{\sigma^2}{\mu_1 - \mu_0} \log \frac{P_Y(0)}{P_Y(1)} \]

- the prior moves the threshold up or down, depending on the probabilities of 0s and 1s.

how relevant is the prior?

- it is weighed by the inverse of the normalized distance between the means

\[ \frac{1}{\sqrt{\frac{\mu_1 - \mu_0}{\sigma^2}}} \]

- if the classes are very far apart, the prior makes no difference
- if the classes are exactly equal (same mean) the prior gets infinite weight
The Gaussian classifier

- this is one example of a Gaussian classifier
  - in practice we rarely have only one variable
  - typically $X = (X_1, \ldots, X_n)$ is a vector of observations

- the BDR for this case is equivalent, but more interesting

- the central different is the class-conditional distributions are multivariate Gaussian

\[
P_{X|Y}(x | i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left\{-\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right\}
\]
The Gaussian classifier

in this case

\[
P_{X|Y}(x \mid i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left\{ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right\}
\]

• the BDR

\[
i^*(X) = \arg \max_i \left[ \log P_{X|Y}(X \mid i) + \log P_Y(i) \right]
\]

• becomes

\[
i^*(X) = \arg \max_i \left[ -\frac{1}{2} (X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i) \\
- \frac{1}{2} \log(2\pi)^d |\Sigma_i| + \log P_Y(i) \right]
\]
The Gaussian classifier

- This can be written as

\[ i^*(x) = \arg \min_i [d_i(x, \mu_i) + \alpha_i] \]

with

\[ d_i(x, y) = (x - y)^T \Sigma_i^{-1} (x - y) \]

\[ \alpha_i = \log(2\pi)^d |\Sigma_i| - 2 \log P_Y(i) \]

- The optimal rule is to assign \( x \) to the closest class.
- Closest is measured with the Mahalanobis distance \( d_i(x,y) \).
- To which the \( \alpha \) constant is added to account for the class prior.
The Gaussian classifier

first special case of interest:
- all classes have the same covariance,

$$\Sigma_i = \Sigma, \quad \forall i$$

the BDR becomes

$$i^*(x) = \arg \min_i [d(x, \mu_i) + \alpha_i]$$

- with

$$d(x, y) = (x - y)^T \Sigma^{-1} (x - y)$$

$$\alpha_i = \log(2\pi)^d |\Sigma| - 2 \log P_Y(i)$$

same metric for all classes
constant, not function of i, can be dropped
The Gaussian classifier

in detail

\[ i^* (x) = \arg \min \left[ (x - \mu_i)^T \Sigma^{-1} (x - \mu_i) - 2 \log P_Y(i) \right] \]

\[ = \arg \min \left[ x^T \Sigma^{-1} x - x^T \Sigma^{-1} \mu_i - \mu_i^T \Sigma^{-1} x + \mu_i^T \Sigma^{-1} \mu_i - 2 \log P_Y(i) \right] \]

\[ = \arg \min \left[ x^T \Sigma^{-1} x - 2 \mu_i^T \Sigma^{-1} x + \mu_i^T \Sigma^{-1} \mu_i - 2 \log P_Y(i) \right] \]

\[ = \arg \max \left[ \mu_i^T \Sigma^{-1} x - \frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P_Y(i) \right] \]
The Gaussian classifier

- in summary,

\[ i^*(x) = \arg \max_i g_i(x) \]

- with

\[ g_i(x) = w_i^T x + w_{i0} \]

\[ w_i = \Sigma_i^{-1} \mu_i \]

\[ w_{i0} = -\frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i + \log P_Y(i) \]

- the BDR is a linear function or a linear discriminant
Geometric interpretation

- classes \(i,j\) share a boundary if
  - there is a set of \(x\) such that
    \[
    g_i(x) = g_j(x)
    \]
  - or
    \[
    \left( w_i - w_j \right)^T x + \left( w_{i0} - w_{j0} \right) = 0
    \]
    \[
    \left( \Sigma^{-1} \mu_i - \Sigma^{-1} \mu_j \right)^T x + \left( -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P_Y(i) + \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j - \log P_Y(j) \right) = 0
    \]
Geometric interpretation

note that

\[
\begin{align*}
\left(\Sigma^{-1}\mu_i - \Sigma^{-1}\mu_j\right)^T x + \\
\left(-\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P_Y(i) + \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j - \log P_Y(j)\right) = 0
\end{align*}
\]

• can be written as

\[
\begin{align*}
\left(\mu_i - \mu_j\right)^T \Sigma^{-1} x - \frac{1}{2} \left(\mu_i^T \Sigma^{-1} \mu_i - \mu_j^T \Sigma^{-1} \mu_j - 2 \log \frac{P_Y(i)}{P_Y(j)}\right) = 0
\end{align*}
\]

next, we use

\[
\begin{align*}
\mu_i^T \Sigma^{-1} \mu_i - \mu_j^T \Sigma^{-1} \mu_j = \\
\mu_i^T \Sigma^{-1} \mu_i - \mu_i^T \Sigma^{-1} \mu_j + \mu_i^T \Sigma^{-1} \mu_j - \mu_j^T \Sigma^{-1} \mu_j =
\end{align*}
\]
Geometric interpretation

which can be written as

\[
\mu_i^T \Sigma^{-1} \mu_i - \mu_j^T \Sigma^{-1} \mu_j =
\]

\[
\mu_i^T \Sigma^{-1} \mu_i - \mu_i^T \Sigma^{-1} \mu_j + \mu_i^T \Sigma^{-1} \mu_j - \mu_j^T \Sigma^{-1} \mu_j =
\]

\[
\mu_i^T \Sigma^{-1} (\mu_i - \mu_j) + (\mu_i - \mu_j)^T \Sigma^{-1} \mu_j =
\]

\[
\mu_i^T \Sigma^{-1} (\mu_i - \mu_j) + \mu_j^T \Sigma^{-1} (\mu_i - \mu_j) =
\]

\[
(\mu_i + \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)
\]

using this in

\[
(\mu_i - \mu_j)^T \Sigma^{-1} x - \frac{1}{2} \left( \mu_i^T \Sigma^{-1} \mu_i - \mu_j^T \Sigma^{-1} \mu_j - 2 \log \frac{P_Y(i)}{P_Y(j)} + \right) = 0
\]
Geometric interpretation

leads to

\[
\left(\mu_i - \mu_j\right)^T \Sigma^{-1} x - \frac{1}{2} \left(\left(\mu_i + \mu_j\right)^T \Sigma^{-1} (\mu_i - \mu_j) - 2 \log \frac{P_Y(i)}{P_Y(j)} + \right) = 0
\]

\[
w^T x + b = 0
\]

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

\[
b = -\frac{\left(\mu_i + \mu_j\right)^T \Sigma^{-1} (\mu_i - \mu_j)}{2} + \log \frac{P_Y(i)}{P_Y(j)}
\]

this is the equation of the hyper-plane of parameters \(w\) and \(b\)
Geometric interpretation

which can also be written as

\[
(\mu_i - \mu_j)^T \Sigma^{-1} x - \frac{1}{2} \left( (\mu_i + \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j) - 2 \log \frac{P_Y(i)}{P_Y(j)} \right) = 0
\]

\[
(\mu_i - \mu_j)^T \Sigma^{-1} \left( x - \frac{\mu_i + \mu_j}{2} + \frac{\mu_i - \mu_j}{(\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)} \log \frac{P_Y(i)}{P_Y(j)} \right) = 0
\]

or

\[
W^T (x - x_0) = 0
\]

\[
W = \Sigma^{-1} (\mu_i - \mu_j)
\]

\[
x_0 = \frac{\mu_i + \mu_j}{2} - \frac{(\mu_i - \mu_j)}{(\mu_i - \mu_j)^T \Sigma^{-1} (\mu_i - \mu_j)} \log \frac{P_Y(i)}{P_Y(j)}
\]
Geometric interpretation

- this is the equation of the hyper-plane
  - of normal vector $w$
  - that passes through $x_0$

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

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

$$x_0 = \frac{\mu_i + \mu_j}{2} - \frac{(\mu_i - \mu_j) \Sigma^{-1}(\mu_i - \mu_j)}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)} \log \frac{P_Y(i)}{P_Y(j)}$$

optimal decision boundary for Gaussian classes, equal covariance
Geometric interpretation

▶ special case i)

\[ \Sigma = \sigma^2 I \]

▶ optimal boundary has

\[
W = \frac{\mu_i - \mu_j}{\sigma^2} \\
X_0 = \frac{\mu_i + \mu_j}{2} - \sigma^2 \frac{(\mu_i - \mu_j)}{\|\mu_i - \mu_j\|^2} \log \frac{P_Y(i)}{P_Y(j)} \\
= \frac{\mu_i + \mu_j}{2} - \sigma^2 \frac{1}{\|\mu_i - \mu_j\|^2} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j)
\]
Geometric interpretation

this is

\[ W = \frac{\mu_i - \mu_j}{\sigma^2} \]

\[ \chi_0 = \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j) \]

vector along the line through \( \mu_i \) and \( \mu_j \)

Gaussian classes, equal covariance \( \sigma^2 I \)
Geometric interpretation

- for equal prior probabilities \( (P_Y(i) = P_Y(j)) \)
  - optimal boundary:
    - plane through midpoint between \( \mu_i \) and \( \mu_j \)
    - orthogonal to the line that joins \( \mu_i \) and \( \mu_j \)

\[
W = \frac{\mu_i - \mu_j}{\sigma^2}
\]
\[
X_0 = \frac{\mu_i + \mu_j}{2}
\]

mid-point between \( \mu_i \) and \( \mu_j \)

Gaussian classes, equal covariance \( \sigma^2 I \)
Geometric interpretation

- different prior probabilities \( (P_Y(i) \neq P_Y(j)) \)

\[
\mathcal{W} = \frac{\mu_i - \mu_j}{\sigma^2}
\]
\[
\chi_0 = \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j)
\]

\( x_0 \) moves along line through \( \mu_i \) and \( \mu_j \)

Gaussian classes, equal covariance \( \sigma^2 I \)
Geometric interpretation

what is the effect of the prior? \( (P_Y(i) \neq P_Y(j) ) \)

\[
x_0 = \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j)
\]

\( x_0 \) moves away from \( \mu_i \) if \( P_Y(i) > P_Y(j) \)
making it more likely to pick \( i \).

Gaussian classes, equal covariance \( \sigma^2 I \)
Geometric interpretation

what is the strength of this effect? \((P_Y(i) \neq P_Y(j))\)

\[
W = \frac{\mu_i - \mu_j}{\sigma^2}
\]

\[
\chi_0 = \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j)
\]

“inversely proportional to the distance between means in units of standard deviation”

Gaussian classes, equal covariance \(\sigma^2 I\)
Geometric interpretation

- note the similarities with scalar case, where

\[ X < \frac{\mu_i + \mu_j}{2} + \frac{\sigma^2}{\mu_i - \mu_j} \log \frac{P_Y(0)}{P_Y(1)} \]

- while here we have

\[ \mathbf{W}^T (X - X_0) = 0 \]
\[ \mathbf{W} = \frac{\mu_i - \mu_j}{\sigma^2} \]
\[ X_0 = \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j) \]

- hyper-plane is the high-dimensional version of the threshold!
Geometric interpretation

- boundary hyper-plane in 1, 2, and 3D
- for various prior configurations
Geometric interpretation

special case ii) \[ \Sigma_i = \Sigma \]

optimal boundary

\[
\mathbf{w}^T (\mathbf{x} - \mathbf{x}_0) = 0
\]
\[
\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)
\]
\[
\mathbf{x}_0 = \frac{\mu_i + \mu_j}{2} - \frac{1}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)} \log \frac{P_Y(i)}{P_Y(j)} (\mu_i - \mu_j)
\]

- \(x_0\) basically the same, strength of the prior inversely proportional to Mahalanobis distance between means
- \(w\) is multiplied by \(\Sigma^{-1}\), which changes its direction and the slope of the hyper-plane
Geometric interpretation

- **equal but arbitrary covariance**

\[
\mathbf{w} = \Sigma^{-1}(\mathbf{\mu}_i - \mathbf{\mu}_j)
\]

\[
\chi_0 = \frac{\mathbf{\mu}_i + \mathbf{\mu}_j}{2} - \frac{1}{(\mathbf{\mu}_i - \mathbf{\mu}_j)^T \Sigma^{-1}(\mathbf{\mu}_i - \mathbf{\mu}_j)} \log \frac{P_Y(i)}{P_Y(j)} (\mathbf{\mu}_i - \mathbf{\mu}_j)
\]

Gaussian classes, equal covariance \( \Sigma \)
Geometric interpretation

In the homework you will show that the separating plane is tangent to the pdf iso-contours at $x_0$.

- Gaussian classes, equal covariance $\Sigma$
  - reflects the fact that the natural distance is now Mahalanobis
Geometric interpretation

- boundary hyper-plane in 1, 2, and 3D
- for various prior configurations
Geometric interpretation

- what about the generic case where covariances are different?
  - in this case
    \[
    i^*(x) = \arg \min_i [d_i(x, \mu_i) + \alpha_i]
    \]
    \[
    d_i(x, y) = (x - y)^T \Sigma_i^{-1} (x - y)
    \]
    \[
    \alpha_i = \log(2\pi)^d |\Sigma_i| - 2\log P_Y(i)
    \]
  - there is not much to simplify
    \[
    g_i(x) = (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) + \log |\Sigma_i| - 2\log P_Y(i)
    \]
    \[
    = x^T \Sigma_i^{-1} x - 2x^T \Sigma_i^{-1} \mu_i + \mu_i^T \Sigma_i^{-1} \mu_i + \log |\Sigma_i| - 2\log P_Y(i)
    \]
Geometric interpretation

\[ g_j(x) = x^T \Sigma_j^{-1} x - 2x^T \Sigma_i^{-1} \mu_j + \mu_j^T \Sigma_j^{-1} \mu_j + \log |\Sigma_j| - 2 \log P_Y(i) \]

• which can be written as

\[ g_j(x) = x^T \mathcal{W}_j x + \mathcal{W}_j^T x + \mathcal{W}_{j0} \]

\[ \mathcal{W}_j = \Sigma_j^{-1} \]

\[ \mathcal{W}_j = -2 \Sigma_j^{-1} \mu_j \]

\[ \mathcal{W}_{j0} = \mu_j^T \Sigma_j^{-1} \mu_j + \log |\Sigma_j| - 2 \log P_Y(i) \]

► for 2 classes the decision boundary is hyper-quadratic

• this could mean hyper-plane, pair of hyper-planes, hyper-spheres, hyper-ellipsoids, hyper-hyperboloids, etc.
Geometric interpretation

in 2 and 3D:
The sigmoid

we have derived all of this from the log-based BDR

\[ i^* (X) = \arg \max_i \left[ \log P_{X|Y} (X \mid i) + \log P_Y (i) \right] \]

when there are only two classes, it is also interesting to look at the original definition

\[ i^* (X) = \arg \max_i g_i (X) \]

with

\[ g_i (X) = P_{Y|X} (i \mid X) = \frac{P_{X|Y} (X \mid i) P_Y (i)}{P_X (X)} \]

\[ = \frac{P_{X|Y} (X \mid i) P_Y (i)}{P_{X|Y} (X \mid 0) P_Y (0) + P_{X|Y} (X \mid 1) P_Y (1)} \]
The sigmoid

- note that this can be written as
  \[ i^*(x) = \arg \max \ g_i(x) \]
  \[ g_i(x) = 1 - g_0(x) \]

- and, for Gaussian classes, the posterior probabilities are
  \[ g_0(x) = \frac{1}{1 + \exp\{d_0(x - \mu_0) - d_1(x - \mu_1) + \alpha_0 - \alpha_1\}} \]

- where, as before,
  \[ d_i(x, y) = (x - y)^T \Sigma_i^{-1} (x - y) \]
  \[ \alpha_i = \log(2\pi)^d |\Sigma_i| - 2 \log P_Y(i) \]
The sigmoid

The posterior

$$g_0(x) = \frac{1}{1 + \exp\{d_0(x - \mu_0) - d_1(x - \mu_1) + \alpha_0 - \alpha_1\}}$$

is a sigmoid and looks like this

discriminant: $P(C_1|x) = 0.5$
The sigmoid

- **the sigmoid appears in neural networks**
  - it is the true posterior for Gaussian problems where the covariances are the same

![Graph of Equal variances](image1)

![Graph of Single boundary at halfway between means](image2)
The sigmoid

but not necessarily when the covariances are different

Variances are different

Two boundaries
Bayesian decision theory

advantages:

• BDR is optimal and cannot be beaten
• Bayes keeps you honest
• models reflect causal interpretation of the problem, this is how we think
• natural decomposition into “what we knew already” (prior) and “what data tells us” (CCD)
• no need for heuristics to combine these two sources of info
• BDR is, almost invariably, intuitive
• Bayes rule, chain rule, and marginalization enable modularity, and scalability to very complicated models and problems

problems:

• BDR is optimal only insofar the models are correct.
Any questions?