Bayes Decision Theory and Parameter Estimation

Ulf Brefeld and Klaus R. Müller

Machine Learning Group

Overview

1. Bayes Decision Theory
   - Bayes' Decision Rule
   - Risk Minimization
   - Discriminant Functions and Decision Boundaries
   - Discriminants for Gaussian Distributions

2. Parameter Estimation
   - Maximum Likelihood
   - Example

An arbitrary *classification* task:

- Input/observation: *feature vector* $\mathbf{x} \in \mathcal{X}$.
  - $\mathbf{x}$ is an abstraction of a real-world object.
  - Frequently, the input space is $\mathcal{X} = \mathbb{R}^d$
- Output/explanation: random variable of interest $y \in \mathcal{Y}$.
  - Binary classification $\mathcal{Y} = \{+1, -1\}$.
  - Multi-class classification $\mathcal{Y} = \{1, 2, 3, \ldots, k\}$.

Assumption: We know the joint probability distribution $p(\mathbf{x}, y) = p(\mathbf{x}|y)p(y)$

Approach:

$$P(explanation|observation) \propto P(observation|explanation) \times P(explanation)$$

Example

A particular *classification* task: Classify messages as spam or ham.

- Observations: Messages translated into *feature vectors* $\mathbf{x} \in \mathbb{B}^d$.
  - E.g., $\mathbf{x} = (x_1, x_2, \ldots, x_d)'$ may be a bag-of-words encoding.
  - x_1: occurence of the word *Aachen*
  - x_2: occurence of the word *Aar*

    $\vdots$
  - x_d: occurence of the word *ZZ-TOP*
- Class labels: $\mathcal{Y} = \{+1, -1\}$.
  - $+1$: instance is spam
  - -1: instance is ham

Assumption: We know the join probability distribution $p(\mathbf{x}, y) = p(\mathbf{x}|y)p(y)$

Approach:

$$P(\mathit{explanation}|\mathit{observation}) \propto P(\mathit{observation}|\mathit{explanation}) \times P(\mathit{explanation})$$

Bayes' Theorem

$$P(explanation|observation) \propto P(observation|explanation) \times P(explanation)$$

Bayes' Theorem:

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}$$

We call ...

- $P(y|\mathbf{x})$ the *posterior* probability,
- $P(\mathbf{x}|y)$ the *likelihood*,
- $P(y)$ the *prior* probability, and
- $P(\mathbf{x}) = \sum_{\bar{y} \in \mathcal{Y}} P(\mathbf{x}|\bar{y})P(\bar{y})$ the *evidence*.

Decision Rule

A decision rule is a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ that assigns each input $\mathbf{x} \in \mathcal{X}$ to a class label $y \in \mathcal{Y}$.

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}$$

A decision $\hat{y}$ for a given $\mathbf{x}$ is incorrect if $y_{true} \neq \hat{y}$. In a 2-class scenario we have,

$$P(error|\mathbf{x}) = \begin{cases} P(+1|\mathbf{x}) : y_{true} = -1 \\ P(-1|\mathbf{x}) : y_{true} = +1, \end{cases}$$

Obviously, it holds: $P(correct|\mathbf{x}) = 1 - P(error|\mathbf{x})$.

The Bayes' Decision Rule

How do we find a *good* decision rule, that is, one that minimizes the expected error?

$$P(error) = \int_{\mathbf{x} \in \mathcal{X}} P(error|\mathbf{x}) p(\mathbf{x}) \, d\mathbf{x}.$$

If $P(error|\mathbf{x})$ is as small as possible for every $\mathbf{x}$ $\Rightarrow$ the integral must be as small as possible!

The Bayes decision minimizes $P(error)$ and can simply be stated as:

$$\hat{y} = \text{argmax}_{y \in \mathcal{Y}} P(y|\mathbf{x})$$

In other words: Decide in favor of the most probable class!

Bayes' Rule and Decision Boundaries

Bayes' decision rule:

$$f^{Bayes}(\mathbf{x}) = \text{argmax}_{y \in \mathcal{Y}} P(y|\mathbf{x})$$

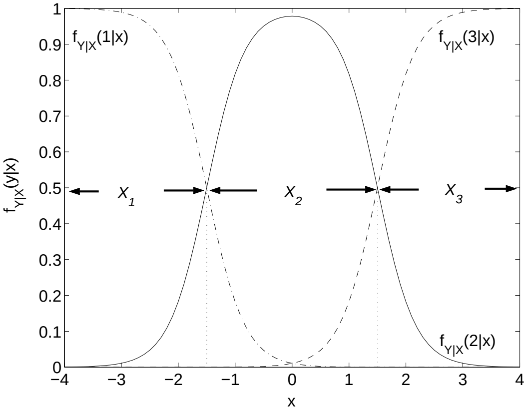
The Bayes decision f^{Bayes} induces regions X_y in input space, associated with class labels,

$$X_y = \{\mathbf{x} : f^{Bayes}(\mathbf{x}) = y\}$$

The decision boundary between classes y and y' is given by the set

$$B_{y,y'} = \{\mathbf{x} : P(y|\mathbf{x}) = P(y'|\mathbf{x})\}, \quad \forall y, y' \in \mathcal{Y}$$

Bayes Decision Rule in Practice



Example with 3 classes.

Problem-dependent Misclassification Costs

Until now, confusing classes cause constant errors irrespectively of the involved classes.

Sometimes, certain errors are more severe than others.

- $\mathcal{Y} = \{AlleKontrollleuchtenImGruenenBereich, Kernschmelze\}$
- $\mathcal{Y} = \{gesund, krank\}$
- $\mathcal{Y} = \{spam, ham\}$

Solution: Introduce loss (or cost) function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \{\mathbb{R}^+ \cup 0\}$.
Example:

	$\hat{y} = spam$	$\hat{y} = ham$
$y_{true} = spam$	$\ell(H, H) = 0$	$\ell(S, H) = 1$
$y_{true} = ham$	$\ell(H, S) = 1000$	$\ell(S, S) = 0$

Drawback: How to define ℓ for a problem at hand?

Example for 2-classes

Define the class-based risks:

$$r(+1,\mathbf{x}) = \ell(+1,+1)P(+1|\mathbf{x}) + \ell(+1,-1)P(-1|\mathbf{x})$$
$$r(-1,\mathbf{x}) = \ell(-1,+1)P(+1|\mathbf{x}) + \ell(-1,-1)P(-1|\mathbf{x})$$

Decide for class $+1$ if $r(+1,\mathbf{x}) < r(-1,\mathbf{x})$, that is,

$$\Big(\ell(-1,+1) - \ell(+1,+1)\Big)P(+1|\mathbf{x}) > \Big(\ell(+1,-1) - \ell(-1,-1)\Big)P(-1|\mathbf{x})$$

For the 0/1-loss, defined as $\ell(a,b) = 1$ if $a \neq b$ and 0 otherwise, we resolve the minimum-error decision: Decide class $+1$ if

$$P(+1|\mathbf{x}) > P(-1|\mathbf{x}).$$

Discriminants for Gaussian Distributed Classes

Bayes' decision rule relies on

$$P(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)P(y)}{\sum_{y' \in \mathcal{Y}} p(\mathbf{x}|y')P(y')}$$

Recall that a minimum-error classification can also be achieved by

$$f_y(\mathbf{x}) = \log p(\mathbf{x}|y) + \log P(y).$$

Let $p(\mathbf{x}|y) \sim N(\boldsymbol{\mu}_y, \Sigma_y)$.

$$f_y(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)'\Sigma_y^{-1}(\mathbf{x} - \boldsymbol{\mu}_y) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_y| + \log P(y)$$

Case 1: Independent Features

$$f_y(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_y)'\Sigma_y^{-1}(\mathbf{x}-\boldsymbol{\mu}_y) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_y| + \log P(y)$$

Consider the simple case $\Sigma_y = \sigma^2\mathbf{1}$ for all $y \in \mathcal{Y}$:

- Features are statistically independent and have the same variance.
- Equal sized hyperspherical clusters centered around the $\boldsymbol{\mu}_y$.

$\Rightarrow$ Determinant $|\Sigma| = \sigma^{2d}$, inverse $\Sigma^{-1} = (1/\sigma^2)\mathbf{1}$

Obtain linear discriminant function:

$$\begin{aligned}
f_y(\mathbf{x}) &= -\frac{1}{2\sigma^2}\|\mathbf{x}-\boldsymbol{\mu}_y\|^2 + \log P(y) \\
&= -\frac{1}{2\sigma^2}[\mathbf{x}'\mathbf{x} - 2\boldsymbol{\mu}_y'\mathbf{x} + \boldsymbol{\mu}_y'\boldsymbol{\mu}_y] + \log P(y) \\
&= \underbrace{\frac{1}{\sigma^2}\boldsymbol{\mu}_y'}_{=:\mathbf{w}_y}\mathbf{x} + \underbrace{\left(-\frac{1}{2\sigma^2}\boldsymbol{\mu}_y'\boldsymbol{\mu}_y + \log P(y)\right)}_{=:b_y} \\
&= \mathbf{w}_y'\mathbf{x} + b_y
\end{aligned}$$

Case 2: Identical Covariance Matrices

$$f_y(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)'\Sigma_y^{-1}(\mathbf{x} - \boldsymbol{\mu}_y) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_y| + \log P(y)$$

Consider the simple case $\Sigma_y = \Sigma$ for all $y \in \mathcal{Y}$:

- Hyperellipsoidal clusters of equal size and shape, centered around the $\boldsymbol{\mu}_y$.

$\Rightarrow$ Again, $|\Sigma|$ and $(d/2)\log 2\pi$ can be ignored.

Obtain linear discriminant function:

$$\begin{aligned} f_y(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_y) + \log P(y) \\ &= -\frac{1}{2}[\mathbf{x}'\Sigma^{-1}\mathbf{x} - 2\boldsymbol{\mu}_y'\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}_y'\Sigma^{-1}\boldsymbol{\mu}_y] + \log P(y) \\ &= \underbrace{\Sigma^{-1}\boldsymbol{\mu}_y'}_{=:\mathbf{w}_y}\mathbf{x} + \underbrace{\left(-\frac{1}{2}\boldsymbol{\mu}_y'\Sigma^{-1}\boldsymbol{\mu}_y + \log P(y)\right)}_{=:b_y} \\ &= \mathbf{w}_y'\mathbf{x} + b_y \end{aligned}$$

Resulting Decision Rule

$$f_y(\mathbf{x}) = \mathbf{w}_y'\mathbf{x} + b_y, \quad \forall y \in \mathcal{Y}$$

Compute the resulting decision as follows.

- Multi-class case, $\mathcal{Y} = \{1, 2, 3, \ldots, k\}$:

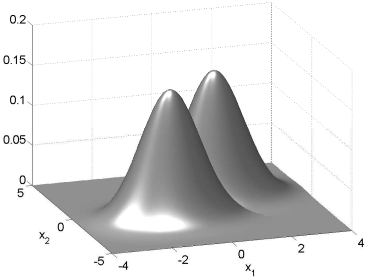
$$\hat{y} = f(\mathbf{x}) = \mathrm{argmax}_{y \in \mathcal{Y}} f_y(\mathbf{x})$$

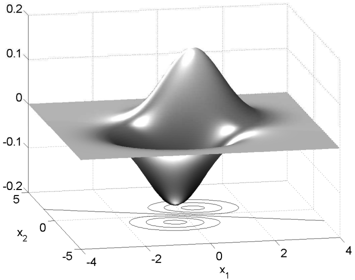
- Binary case, $\mathcal{Y} = \{+1, -1\}$, decide $+1$ if

$$\begin{aligned}
f_{+1}(\mathbf{x}) &> f_{-1}(\mathbf{x}) \\
\mathbf{w}_{+1}'\mathbf{x} + b_{+1} &> \mathbf{w}_{-1}'\mathbf{x} + b_{-1} \\
\mathbf{w}_{+1}'\mathbf{x} + b_{+1} - \mathbf{w}_{-1}'\mathbf{x} - b_{-1} &> 0 \\
\underbrace{(\mathbf{w}_{+1} - \mathbf{w}_{-1})}_{=:\mathbf{w}}{}'\mathbf{x} + \underbrace{(b_{+1} - b_{-1})}_{=:b} &> 0 \\
\mathbf{w}'\mathbf{x} + b &> 0,
\end{aligned}$$

  and -1 otherwise.

Linear Discriminant Functions





Left: Posterior class distribution. Right: Decision boundary.

Parameter Estimation

Parameter Estimation

Recall: $p(\mathbf{x}, y) = p(\mathbf{x}|y)P(y)$.

Bayes' decision rule only applicable when $P(y)$ and class-conditional densities $p(\mathbf{x}|y)$ are known.

In general, $P(y)$ and $p(\mathbf{x}|y)$ are unknown in practical applications!

Instead, we are given a set of (training) samples D drawn *independent and identically distributed (iid)* from $p(\mathbf{x}, y)$.

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$$

Task: Use this sample to estimate $P(y)$ and $p(\mathbf{x}|y)$!

Estimating the Prior

Given: iid training sample of size n,

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$$

Task: Estimate prior $P(y)$. Solve by simply counting:

- How many times have we seen label y in the training set?
- Normalize to obtain probabilities!

$$\hat{P}(y) = \frac{\sum_{i=1}^{n} 1_{[y_i = y]}}{n}, \quad \forall y \in \mathcal{Y}$$

The larger the sample size n, the better will be the estimate $\hat{P}(y)$.

Estimating the Class-conditional Densities

Given: iid training sample of size n,

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$$

Task: Estimate class-conditional $p(\mathbf{x}|y)$.

Difficult for several reasons:

- $\mathcal{X}$ is usually high-dimensional, often $\#dimensions \gg n$
- $\Rightarrow$ density estimation will be poor in sparse regions

$\Rightarrow$ We need assumptions!

- E.g., density to be estimated is Gaussian with unknown $\boldsymbol{\mu}$ and Σ
- Instead of inferring an unknown function $p(\mathbf{x}|y)$ now only parameters need to be estimated!
- $\Rightarrow$ Maximum likelihood!

Given: class-conditional density $p(\mathbf{x}|y;\theta_y)$ in parametric form, iid sample D
Task: Find parameters θ_y such that the likelihood of the data is maximized

Assume further that the θ_y are functionally independent
$\Rightarrow$ Deal with each class separately and simplify notation

For each class y let $D_y = \{\mathbf{x} : (\mathbf{x}, \bar{y}) \in D,\ y = \bar{y}\}$ such that

$$D_y = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$$

Since D (and hence also D_y) are drawn iid, we have

$$p(D_y|\theta_y) = \prod_{i=1}^{m} p(\mathbf{x}_i|\theta_y)$$

Maximize the likelihood $p(D|\theta) = \prod_{i=1}^{m} p(\mathbf{x}_i|\theta)$ by finding parameters $\theta = (\theta_1, \ldots, \theta_q)'$ that *agree* with the data.

Log-likelihood:

$$\log L(\theta) = \log p(D_y|\theta) = \sum_{i=1}^{m} \log p(\mathbf{x}_i|\theta)$$

Compute partial derivatives

$$\frac{\partial \log L}{\partial \theta_1}, \frac{\partial \log L}{\partial \theta_2}, \ldots, \frac{\partial \log L}{\partial \theta_q},$$

and find optimal θ^* by solving

$$\frac{\partial \log L}{\partial \theta_1} \stackrel{!}{=} 0, \frac{\partial \log L}{\partial \theta_2} \stackrel{!}{=} 0, \ldots, \frac{\partial \log L}{\partial \theta_q} \stackrel{!}{=} 0.$$

ML Example: Multivariate Normal

Example:

$$\log p(\mathbf{x}_i|\boldsymbol{\mu},\Sigma) = -\frac{1}{2}\log\{(2\pi)^d|\Sigma|\} - \frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})$$

Differentiate wrt μ:

$$\frac{\partial \log p(\mathbf{x}_i|\boldsymbol{\mu},\Sigma)}{\partial \boldsymbol{\mu}} = \Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})$$

$\Rightarrow$ The optimal $\hat{\boldsymbol{\mu}}$ must satisfy

$$\sum_{i=1}^{m}\Sigma^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}) = 0$$

Multiplying with Σ and re-arranging the terms leads to

$$\hat{\boldsymbol{\mu}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$$