Probability Models for Bayesian Recognition

Notation .................................................................2

Supervised Learning for Bayesian Classification ..............3
  Ratio of Histograms (Recall) ........................................4

Variable Sized Histogram Cells..........................................5

Kernel Density Estimators .............................................6

K Nearest Neighbors .....................................................8

Probability Density Functions .........................................9

The Normal Density Function ......................................10
  Univariate Normal ................................................10
  Multivariate Normal ...............................................10
  The Mahalanobis Distance .......................................11
  Expected Values and Moments ................................12
  Biased and Unbiased Variance ..................................12

! Bookmark not defined.

Bibliographical sources:
Notation

- $x$: A variable
- $X$: A random variable (unpredictable value), an observation.
- $N$: The number of possible values for $X$
- $\bar{x}$: A vector of $D$ variables.
- $\bar{X}$: A vector of $D$ random variables.
- $D$: The number of dimensions for the vector $\bar{x}$ or $\bar{X}$
- $C_k$: The class $k$
- $k$: Class index
- $K$: Total number of classes
- $\omega_k$: The statement (assertion) that $E \in C_k$
- $P(\omega_k) = P(X \in C_k)$: Probability that the observation $E$ is a member of the class $k$.
- $M_k$: Number of examples for the class $k$.
- $M$: Total number of examples.

$$M = \sum_{k=1}^{K} M_k$$

- $\{\bar{X}_m\}$: A set of training samples
- $\{y_m\}$: A set of indicator vectors for the training samples in $\{\bar{X}_m\}$
- $p(X)$: Probability density function for a continuous value $X$
- $p(\bar{X})$: Probability density function for continuous $\bar{X}$
- $p(\bar{X} \mid \omega_k)$: Probability density for $\bar{X}$, the class $k$. $\omega_k = X \in C_k$.
- $Q$: Number of cells in $h(x)$. $Q = ND$
- $S$: A sum of $V$ adjacent histogram cells: $S = \sum_{\bar{x} \in V} h(\bar{x})$
- $V$: The "Volume" of the region of the histogram
Supervised Learning for Bayesian Classification

Our problem is to build a box that maps a set of features \( \tilde{X} \) from an observation, \( X \) to a class \( C_k \) from a set of \( K \) possible classes.

![Diagram: \( \tilde{X} \rightarrow \text{Classify} \rightarrow \hat{c} \in \{C_i\} \)]

Let \( \omega_k \) be the proposition that the event belongs to class \( k \): \( \omega_k = \tilde{X} \in C_k \)

In order to minimize the number of mistakes, we will maximize the probability that \( \omega_k \equiv X \in C_k \)

\[
\hat{\omega}_k = \arg\max_{\omega_k} \left\{ P(\omega_k \mid \tilde{X}) \right\} 
\]

Our primary tool for this is Baye's Rule:

\[
P(\omega_k \mid \tilde{X}) = \frac{P(\tilde{X} \mid \omega_k)P(\omega_k)}{P(\tilde{X})} = \frac{P(\tilde{X} \mid \omega_k)P(\omega_k)}{\sum_{k=1}^{K} P(\tilde{X} \mid \omega_k)}
\]

We have already seen how to use histograms to represent \( P(\tilde{X} \mid \omega_k) \), and \( P(\tilde{X}) \) and to use \( P(\omega_k) = \frac{M_k}{M} \)

To apply Baye’s rule, we require a representation for the probabilities \( P(\tilde{X} \mid \omega_k) \), \( P(\tilde{X}) \), and \( p(\omega_k) \).

In the next few lectures we will look at techniques that learn to recognize classes using Bayes rule. These methods are based on an explicit probabilistic model of the training data. We will start with some simple, non-parametric models. We will then look at models using the Gaussian or Normal density functions. This will lead to the use of weighted sums of Gaussians (Gaussian Mixture Models).

Today will look at three non-parametric representations for \( P(\tilde{X} \mid \omega_k) \) and \( P(\tilde{X}) \)

1) Kernel Density Estimators
2) K-Nearest Neighbors
3) Probability density functions (PDF)


**Ratio of Histograms (Recall)**

Consider an example of K classes of objects where objects are described by a feature, $X$, with $N$ possible values from $[1,N]$. Assume that we have a "training set" of $M$ samples $\{X_m\}$ along with indicator variables $\{y_m\}$ where the indicator variable is the class, $k$, for each training sample.

For each class $k$, we allocate a histogram, $h_k()$, with $N$ cells and count the values in the training set.

\[ \forall_{m=1}^{M} : h(X) \leftarrow h(X_m) + 1 \]

if $y_m = k$ THEN $h_k(X_m) \leftarrow h_k(X_m) + 1$; $M_k \leftarrow M_k + 1$

Then

\[ p(X=x) = \frac{1}{M} h(x) \]

\[ p(X=x | X \in C_k) = p(X | \omega_k) = \frac{1}{M_k} h_k(x) \]

and $p(\omega_k)$ can be estimated from the relative size of the training set.

\[ p(X \in C_k) = p(\omega_k) = \frac{M_k}{M} \]

giving:

\[ p(\omega_k | X) = \frac{p(X | \omega_k) p(\omega_k)}{p(X)} = \frac{1}{M} h(X) \frac{M_k}{M} = \frac{h_k(X)}{h(X)} = T(X) \]

This can be represented by a lookup table.

To illustrate, consider an example with 2 classes ($K=2$) and where $X$ can take on 8 values ($N=8$, $D=1$).

![Histogram](image)

Recall that the number of cells in the histogram is $Q=N^D$.

Having $M >> Q$ is NECESSARY but NOT Sufficient.

NOT having $M >> Q$ is a guarantee of INSUFFICIENT TRAINING DATA.

So what can you do if $M$ is not $>> Q$?

Adapt the size of the cell to the data!
Variable Sized Histogram Cells

Suppose that we have a D-dimensional feature vector \( \bar{X} \) with each feature quantized to \( N \) possible values, and suppose that we represent \( p(\bar{X}) \) as a D-dimensional histogram \( h(\bar{X}) \). Let us fill the histogram with M training samples \( \{\bar{X}_m\} \).

Let us define the volume of each cell as 1.
The volume for any block of \( V \) cells is \( V \).
Then the volume of the entire space is \( Q=N^D \).

If the quantity of training data is too small, ie if \( M < 8Q \), then we can combine adjacent cells so as to amass enough data for a reasonable estimate.

Suppose we merge \( V \) adjacent cells such that we obtain a combined sum of \( S \).

\[
S = \sum_{\bar{X} \in V} h(\bar{X})
\]

The volume of the combined cells would be \( V \).
To compute the probability we replace \( h(\bar{X}) \) with \( \frac{S}{V} \).
The probability \( p(\bar{X}) \) for \( \bar{X} \in V \) is:

\[
p(\bar{X} \in V) = \frac{1}{M} \cdot \frac{S}{V}
\]

This is typically written as:

\[
p(\bar{X}) = \frac{S}{MV}
\]

We can use this equation to develop two alternative non-parametric methods.

Fix \( V \) and determine \( S \) \( \Rightarrow \) Kernel density estimator.
Fix \( S \) and determine \( V \) \( \Rightarrow \) K nearest neighbors.

(note that the symbol “K” is often used for the sum the cells.
This conflicts with the use of \( K \) for the number of classes.
Thus we will use the symbol \( S \) for the sum of adjacent cells).
Kernel Density Estimators

For a Kernel density estimator, we represent each training sample with a kernel function \( k(\tilde{X}) \).

Popular Kernel functions include
- a hypercube centered of side \( w \)
- a triangular function with base of \( w \)
- a sphere of radius \( w \)
- a Gaussian of standard deviation \( \sigma \).

We can define the function for the hypercube as

\[
k(\tilde{u}) = \begin{cases} 
1 & \text{if } |u_d| \leq 1/2 \text{ for all } d = 1, ..., D \\
0 & \text{otherwise}
\end{cases}
\]

This is called a Parzen window.
Subtracting a point, \( \tilde{z} \), centers the Parzen window at that point.
Dividing by \( w \) scales the Parzen window to a hyper-cube of side \( w \).

\[
k\left(\frac{\tilde{X} - \tilde{z}}{w}\right) \text{ is a cube of size } w^D \text{ centered at } \tilde{z}.
\]

The \( M \) training samples define \( M \) overlapping Parzen windows.

For an feature value, \( \tilde{x} \), the probability \( p(\tilde{x}) \) is the sum of Parzen windows at \( \tilde{X} \)

\[
S = \sum_{m=1}^{M} k\left(\frac{\tilde{X} - \tilde{X}_m}{w}\right)
\]

The volume of the Parzen window is \( V = w^D \).

Thus the probability \( p(\tilde{X}) = \frac{S}{MV} = \frac{1}{Mw^D} \sum_{m=1}^{M} k\left(\frac{\tilde{X} - \tilde{X}_m}{w}\right) \)

A Parzen window is discontinuous at the boundaries, creating boundary effects.
We can soften this using a triangular function evaluated within the window.
\[ k(\vec{u}) = \begin{cases} 1 - 2\|\vec{u}\| & \text{if } \|\vec{u}\| \leq 1/2 \\ 0 & \text{otherwise} \end{cases} \]

Even better is to use a Gaussian kernel with standard deviation \( \sigma \).

\[ k(\vec{u}) = \frac{1}{(2\pi)^{D/2} \sigma} e^{-\frac{1}{2} \frac{\|\vec{u}\|^2}{\sigma^2}} \]

We can note that the volume (or integral) of \( e^{-\frac{1}{2} \frac{\|\vec{u}\|^2}{\sigma^2}} \) is \( V = (2\pi)^{D/2} \sigma \)

In this case \( \frac{S}{MV} = \frac{1}{M} \sum_{m=1}^{M} k(\vec{X} - \vec{X}_m) \)

This corresponds to placing a Gaussian at each training sample and summing the Tails at \( \vec{X} \).

The probability for a value \( \vec{X} \) is the sum of the Gaussians.

In fact, we can choose any function \( k(\vec{u}) \) as kernel, provided that

\[ k(\vec{u}) \geq 0 \quad \text{and} \quad \int k(\vec{u})d\vec{u} = 1 \]
K Nearest Neighbors

For K nearest neighbors, we hold S constant and vary V. (We have used the symbol S for the number of neighbors, rather than K to avoid confusion with the number of classes).

For each training sample, $\tilde{X}_m$, we construct a tree structure (such as a KD Tree) that allows us to easily find the S nearest neighbors for any point.

To compute $p(\tilde{X})$ we need the volume of a sphere in D dimensions that encloses the nearest S neighbors. Suppose the set of S nearest neighbors is the set $\{X_i\}$.

This is D dimensional sphere of radius $R = \arg\max_{\{X_i\}} \{\|\tilde{X} - X_i\|\}$

$$V = \frac{\pi^{\frac{D}{2}}}{\Gamma\left(\frac{D}{2} + 1\right)} R^D$$

Where $\Gamma(D) = (D-1)!$

For even D this is easy to evaluate

For odd D, use a table to determine $\Gamma\left(\frac{D}{2} + 1\right)$

Then as before: $p(\tilde{X}) = \frac{S}{MV}$
**Probability Density Functions**

A probability density function \( p(X) \), is a function of a continuous variable \( X \) such that

1) \( X \) is a continuous real valued random variable with values between \([-\infty, \infty]\)
2) \( \int_{-\infty}^{\infty} p(X) = 1 \)

Note that \( p(X) \) is NOT a number but a continuous function.

A probability density function defines the relatively likelihood for a specific value of \( X \). Because \( X \) is continuous, the value of \( p(X) \) for a specific \( X \) is infinitely small. To obtain a probability we must integrate over some range of \( X \).

To obtain a probability we must integrate over some range \( V \) of \( X \).

In the case of \( D=1 \), the probability that \( X \) is within the interval \([A, B]\) is

\[
P(X \in [A, B]) = \int_{A}^{B} p(x) \, dx
\]

This integral gives a number that can be used as a probability.

Note that we use upper case \( P(X \in [A, B]) \) to represent a probability value, and lower case \( p(X) \) to represent a probability density function.

Classification using Bayes Rule can use probability density functions

\[
P(\omega_k | X) = \frac{p(X | \omega_k)}{p(X)} \cdot P(\omega_k) = \frac{p(X | \omega_k)}{\sum_{k=1}^{K} p(X | \omega_k)} \cdot P(\omega_k)
\]

Note that the ratio \( \frac{p(X | \omega_k)}{p(X)} \) IS a number, provided that \( p(X) = \sum_{k=1}^{K} p(X | \omega_k) \)

Probability density functions are easily generalized to vectors of random variables. Let \( \bar{X} \in R^D \), be a vector random variables.

A probability density function, \( p(\bar{X}) \), is a function of a vector of continuous variables

1) \( \bar{X} \) is a vector of \( D \) real valued random variables with values between \([-\infty, \infty]\)
2) \( \int_{-\infty}^{\infty} p(\bar{X}) \, d\bar{x} = 1 \)
The Normal Density Function

Univariate Normal
Whenever a random variable is determined by a sequence of independent random events, the outcome will be a Normal or Gaussian density function. This is known as the Central Limit Theorem. The essence of the derivation is that repeated random events are modeled as repeated convolutions of density functions, and for any finite density function will tend asymptotically to a Gaussian (or normal) function.

\[
p(X) = \mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]

The parameters of \(\mathcal{N}(x; \mu, \sigma)\) are the first and second moments, \(\mu\) and \(\sigma\) of the function.

According to the Central Limit theorem, for any real density \(p(X)\):

\[
as M \to \infty \quad p(X)^M \to \mathcal{N}(x; \mu, \sigma)
\]

Multivariate Normal
For a vector of \(D\) features, \(\tilde{X}\), the Normal density has the form:

\[
p(\tilde{X}) = \mathcal{N}(\tilde{X}; \tilde{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2} (\tilde{X} - \tilde{\mu})^T \Sigma^{-1} (\tilde{X} - \tilde{\mu})}
\]

There are 3 parts to \(\mathcal{N}(\tilde{X}; \tilde{\mu}, \Sigma)\):

\[
\frac{1}{(2\pi)^{D/2} \det(\Sigma)^{1/2}}, \quad e, \quad \text{and} \quad d^2 = -\frac{1}{2} (\tilde{X} - \tilde{\mu})^T \Sigma^{-1} (\tilde{X} - \tilde{\mu})
\]
1) \( e = 2.7818281828 \ldots \) Euler's Constant: \( \int e^x \, dx = e^x \). Used to simplify the algebra.

2) The term \((2\pi)^{-\frac{D}{2}} \det(\Sigma)^{-\frac{1}{2}}\) is a normalization factor.

\[
(2\pi)^{-\frac{D}{2}} \det(\Sigma)^{-\frac{1}{2}} = \int \int \ldots \int e^{-\frac{1}{2}(\vec{X} - \bar{\mu})^T \Sigma^{-1}(\vec{X} - \bar{\mu})} \, dx_1 \, dx_2 \ldots dx_D
\]

The determinant, \( \det(\Sigma) \) is an operation that gives the volume of \( \Sigma \).

For \( D=2 \)

\[
\begin{vmatrix} a & b \\ c & d \end{vmatrix} = a \cdot d - b \cdot c
\]

For \( D > 2 \) this continues recursively.

\[
\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a \begin{vmatrix} e & f \\ i & g \end{vmatrix} + b \begin{vmatrix} f & d \\ i & g \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix}
\]

\[ a(e \cdot i - f \cdot g) + b(f \cdot g - d \cdot i) + c(d \cdot h - e \cdot g) \]

3) The Mahalanobis distance. This is where the action is.

**The Mahalanobis Distance**

The exponent of \( \mathcal{N}(\vec{X}; \bar{\mu}, \Sigma) \) is known as the "Mahalanobis Distance" named for a famous Indian statistician.

\[
d^2 = -\frac{1}{2} (\vec{X} - \bar{\mu})^T \Sigma^{-1} (\vec{X} - \bar{\mu})
\]

This is a distance normalized by the covariance. It is positive and 2nd order (quadratic).

The covariance is provides a distance metric that can be used when the features of \( \vec{X} \) have different units (for example with height (m) and weight (kg)). In this case, the features are compared to the standard deviation of a population.
Expected Values and Moments

The average value is the first moment of the samples

A training set of M samples \( \{X_m\} \) can be used to calculate moments

For M samples of a numerical feature value \( \{X_m\} \), the "expected value" \( E\{X\} \) is defined as the average or the mean:

\[
E\{X\} = \frac{1}{M} \sum_{m=1}^{M} X_m
\]

\( \mu_x = E\{X\} \) is the first moment (or center of gravity) of the values of \( \{X_m\} \).
\( \mu_x = E\{X\} \) is also the first moment (or center of gravity) of the resulting pdf.

\[
E\{X\} = \frac{1}{M} \sum_{m=1}^{M} X_m = \int_{-\infty}^{\infty} p(x) \cdot x \, dx
\]

This is also true for histograms.

The mass of the histogram is \( M = \sum_{x=1}^{N} h(x) \)
Which is also the number of samples used to compute \( h(x) \).

The expected value of \( \{X_m\} \) is the 1st moment of \( h(x) \).

\[
\mu = E\{X_m\} = \frac{1}{M} \sum_{m=1}^{M} X_m = \frac{1}{M} \sum_{x=1}^{N} h(x) \cdot x
\]

For D dimensions, the center of gravity is a vector

\[
\vec{\mu} = E\{\vec{X}\} = \frac{1}{M} \sum_{m=1}^{M} \vec{X} = \left( \begin{array}{c}
E\{X_1\} \\
E\{X_2\} \\
\vdots \\
E\{X_D\}
\end{array} \right)
\]

The vector \( \vec{\mu} \) the vector of averages for the components, \( X_d \) of \( \vec{X} \).
It is also center of gravity (first moment) of the normal density function.

\[
\mu_d = E\{X_{d,m}\} = \frac{1}{M} \sum_{m=1}^{M} X_{d,m} = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(x_1, x_2, \ldots, x_D) \cdot x_d \, dx_1, dx_2, \ldots, dx_D
\]
\[ \bar{\mu} = E(\bar{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_D) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{pmatrix} \]

The variance \( \sigma^2 \) is the square of the expected deviation from the average

\[ \sigma^2 = E((X - \mu)^2) = E(X^2) - \mu^2 = E(X^2) - E(X)^2 \]

This is also the second moment of the pdf

\[ \sigma^2 = E((X - \mu)^2) = \frac{1}{M} \sum_{m=1}^{M} (X_m - \mu)^2 = \int_{-\infty}^{\infty} p(x) \cdot (x - \mu)^2 \, dx \]

and the second moment of a histogram for discrete features.

\[ \sigma^2 = E((X - E(X))^2) = \frac{1}{M} \sum_{m=1}^{M} (X_m - \mu)^2 = \frac{1}{M} \sum_{x=1}^{N} h(x) \cdot (x - \mu)^2 \]

For \( D \) dimensions, the second moment is a co-variance matrix composed of \( D^2 \) terms:

\[ \sigma^2 \equiv E((X_i - \mu_i)(X_j - \mu_j)) = \frac{1}{M} \sum_{m=1}^{M} (X_{i,m} - \mu_i)(X_{j,m} - \mu_j) \]

This is often written

\[ \Sigma = E((\bar{X} - E(\bar{X}))(\bar{X} - E(\bar{X}))^T) \]

and gives

\[ \Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1D}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2D}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{D1}^2 & \sigma_{D2}^2 & \cdots & \sigma_{DD}^2 \end{pmatrix} \]

This provides the parameters for

\[ p(\bar{X}) = \mathcal{N}(\bar{X}; \bar{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(\bar{X} - \bar{\mu})^T \Sigma^{-1}(\bar{X} - \bar{\mu})} \]
The result can be visualized by looking at the equi-probably contours.

Ellipses for 99%, 95%, 90%, 75%, 50%, and 20% of the mass

If $x_i$ and $x_j$ are statistically independent, then $\sigma_{ij}^2 = 0$

For positive values of $\sigma_{ij}^2$, $x_i$ and $x_j$ vary together.
For negative values of $\sigma_{ij}^2$, $x_i$ and $x_j$ vary in opposite directions.

For example, consider features $x_1 = \text{height (meters)}$ and $x_2 = \text{weight (kg)}$

In most people height and weight vary together and so $\sigma_{12}^2$ would be positive