Intelligent Systems: Reasoning and Recognition

James L. Crowley

MOSIG M1 Winter Semester 2021
Lesson 4 11 February 2021

Non-Parametric Models for Bayesian Recognition

Bibliographical sources:

"Pattern Recognition and Machine Learning", C. M. Bishop, Springer Verlag, 2006.
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

Notation

x	A variable
X	A random variable (unpredictable value). an observation.
N	The number of possible values for X
$\vec{x}$	A vector of D variables.
$\vec{X}$	A vector of D random variables.
D	The number of dimensions for the vector $\vec{x}$ or $\vec{X}$
C_k	The class k
k	Class index
K	Total number of classes
ω_k	The statement (assertion) that $X \in C_k$
$P(\omega_k) = P(X \in C_k)$	Probability that the observation X is a member of the class k.
M_k	Number of examples for the class k.
M	Total number of examples. $M = \sum_{k=1}^{K} M_k$
$\{\vec{x}_m\}$	A set of training samples
$\{y_m\}$	A set of indicator vectors for the training samples in $\{\vec{X}_m\}$
$p(X)$	Probability density function for a continuous value X
$p(\vec{X})$	Probability density function for continuous $\vec{X}$
$p(\vec{X} \mid \omega_k)$	Probability density for $\vec{X}$ give the class k. $\omega_k = X \in C_k$.
Q	Number of cells in $h(x)$. $Q = N^D$
S	A sum of V adjacent histogram cells: $S = \sum_{\vec{x} \in V} h(\vec{x})$
V	The "Volume" of the region of the histogram

Bayesian Classification

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

$$\vec{X} \rightarrow \boxed{D(\vec{X})} \rightarrow \hat{C}_k$$

Let ω_k be the proposition that the event belongs to class k: $\omega_k = \vec{X} \in C_k$

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

$$\hat{\omega}_k = \arg\max_{\omega_k} \left\{ P(\omega_k \mid \vec{X}) \right\}$$

Our primary tool for this is Bayes Rule: $P(\omega_k \mid \vec{X}) = \frac{p(\vec{X} \mid \omega_k) P(\omega_k)}{p(\vec{X})}$

To apply Bayes rule, we require a representation for the probabilities $P(\vec{X} \mid \omega_k)$, $P(\vec{X})$, and $P(\omega_k)$. Today we will look at some simple, non-parametric models for probability.

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

1) Histograms
2) Kernel Density Estimators
3) K-Nearest Neighbors

Note that in these cases, P is a capital letter because P() a probability (not a density).

Classification with a Ratio of Histograms

Consider an example of K classes of objects where objects are described by a feature, X, with N possible integer 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.

$$\forall_{m=1}^{M}: \; h(x_m) \leftarrow h(x_m)+1$$
$$\text{IF } y_m = k \text{ THEN } h_k(x_m) \leftarrow h_k(x_m)+1;\; M_k \leftarrow M_k+1$$

Then

$$P(X=x) = \frac{1}{M} h(x)$$

$$P(X=x \mid X \in C_k) = P(X \mid \omega_k) = \frac{1}{M_k} h_k(x)$$

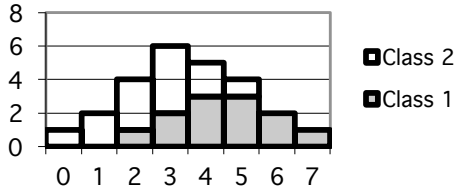
and $P(\omega_k)$ can be estimated from the relative size of the training set.

$$P(X \in C_k) = P(\omega_k) = \frac{M_k}{M}$$

giving: $P(\omega_k \mid X) = \dfrac{P(X \mid \omega_k)P(\omega_k)}{P(X)} = \dfrac{\frac{1}{M_k} h_k(X) \frac{M_k}{M}}{\frac{1}{M} h(X)} = \dfrac{h_k(X)}{h(X)}$

This can also be written as: $P(\omega_k \mid X) = \dfrac{h_k(X)}{\sum_{k=1}^{K} h_k(X)}$ because $h(X) = \sum_{k=1}^{K} h_k(X)$

The ratio of histograms can be represented by a lookup table. $P(\omega_k \mid X) = T(X)$
To illustrate, consider an example with 2 classes (K=2) and where X can take on 8 values (N=8, D=1).



How many training samples are required? How reliable is the result as a function of sample error?

Number of samples required

Problem: Given a feature x, with N possible values, how many observations, M, do we need for a histogram, $h(x)$, to provide a reliable estimate of probability?

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

A common rule is that the Root Mean Square error is proportional to $O(\frac{Q}{M})$.

This can be estimated by comparing the observed histograms to an ideal parametric model of the probability density or by comparing histograms of subsets samples to histograms from a very large sample.

Let $p(x)$ be a probability density function. The RMS (root-mean-square) sampling error between a histogram and the density function is

$$E_{RMS} = \sqrt{E\left\{\left(h(x) - p(x)\right)^2\right\}} \approx O(\frac{Q}{M})$$

However, the constant of proportionality depends on the distribution of data. In the absence of domain knowledge, it is common to assume that all values are equally likely (uniform distribution for x). In the case where p(x) is approximately uniform, $M \geq 8\,Q$ (8 samples per "cell") is reasonable (less than 12% RMS error).

In the non-uniform case, the error is specific to each cell, and we need to guarantee that each cell has received at least 8 samples. This can require a LOT of data.

Having $M >> Q$ is NECESSARY but NOT Sufficient.
Having $M < Q$ is a guarantee of INSUFFICIENT TRAINING DATA.

So what can you do if you do not have $M >> Q$?
Adapt the size of the cell to the data!

Variable Sized Histogram Cells

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

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.

$$S = \sum_{\vec{x} \in V} h(\vec{x})$$

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

$$p(\vec{X} \in V) = \frac{1}{M} \cdot \frac{S}{V}$$

This is typically written as: $p(\vec{X}) = \frac{S}{MV}$

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

Fix V and determine S => Kernel density estimator.
Fix S and determine V => 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 $k(\vec{X})$.

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 σ.

We can define the function for the hypercube as

$$k(\vec{u}) = \begin{cases} 1 & if \ |u_d| \le 1/2 \ \text{ for all } d = 1,...,D \\ 0 & otherwise \end{cases}$$

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

$k\left(\frac{\vec{X}-\vec{z}}{w}\right)$ is a cube of size w^D centered at $\vec{z}$.

We can use the training samples to define the center points.
The M training samples define M overlapping Parzen windows.

$$k\left(\frac{\vec{X}-\vec{x}_m}{w}\right)$$

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

$$S = \sum_{m=1}^{M} k\left(\frac{\vec{X}-\vec{x}_m}{w}\right)$$

The volume of the Parzen window is $V = w^D$.

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

A Parzen window is discontinuous at the boundaries, creating boundary effects.
We can soften this using a triangular function evaluated within the window.

For D=1, the triangular function is:

$$k(u)=\begin{cases} 1-2|\mathrm{u}| & if \quad |\mathrm{u}|\le 1/2 \ \text{ for all d=1,...,D} \\ 0 & otherwise \end{cases}$$

The volume of the triangle of base is w/2 so the probability for an observation x given M training samples $\{X_m\}$ is

$$P(X)=\frac{S}{MV}=\frac{2}{Mw}\sum_{m=1}^{M}k\left(\frac{X-X_m}{w}\right)$$

The Triangular function can be generalized to a D dimensional pyramid using a Manhattan distance.

$$k(\vec{u})=\begin{cases} 1-2|\mathrm{u_d}| & if \quad |\mathrm{u_d}|\le 1/2 \ \text{ for all d=1,...,D} \\ 0 & otherwise \end{cases}$$

It is also possible to generalize the triangle as a hyper-cone with a Euclidean distance

$$k(\vec{u})=\begin{cases} 1-2\|\vec{u}\| & if \quad \|\vec{u}\|\le 1/2 \\ 0 & otherwise \end{cases}$$

Even better is to use a Gaussian kernel with standard deviation σ.

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

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 $P(\vec{X})=\frac{S}{MV}=\frac{1}{M}\sum_{m=1}^{M}k\left(\vec{X}-\vec{x}_m\right)$

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

$$k(\vec{u}) \geq 0 \text{ and } \int k(\vec{u})d\vec{u} = 1$$

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, $\vec{x}_m$, 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 $P(\vec{X})$ 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 $\{X_s\}$.

This is D dimensional sphere of radius $R = \arg\min_{\{X_s\}} \left\{ \left\| \vec{X} - \vec{x}_s \right\| \right\}$

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

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

For even numbered D, $\Gamma(x)$ is easy to evaluate

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

Then as before: $P(\vec{X}) = \frac{S}{MV}$

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 [–∞, ∞]
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

$$P(X \in [A,B]) = \frac{1}{(B-A)} \int_{A}^{B} p(x)\,dx$$

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

$$P(\omega_k \mid X) = \frac{p(X \mid \omega_k)}{p(X)} P(\omega_k) = \frac{p(X \mid \omega_k)}{\sum_{k=1}^{K} p(X \mid \omega_k)}$$

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

Probability density functions are easily generalized to vectors of random variables.
Let $\vec{X} \in R^D$, be a vector random variables.
A probability density function, $p(\vec{X})$, is a function of a vector of continuous variables
1) $\vec{X}$ is a vector of D real valued random variables with values between [–∞, ∞]
2) $\int_{-\infty}^{\infty} p(\vec{x})\,d\vec{x} = 1$

Bayes Rule with probability density functions

Let ω_k represent the statement that a random variable is a member of class C_k: $\omega_k = X \in C_k$. Bayes Rule can be used to compute this probability as:

$$P(\omega_k \mid X) = \frac{p(X \mid \omega_k)}{p(X)} P(\omega_k) = \frac{p(X \mid \omega_k)P(\omega_k)}{\sum_{j=1}^{K} p(X \mid \omega_j)P(\omega_j)}$$

$\frac{p(X \mid \omega_k)}{p(X)}$ IS a number, provided that $p(X) = \sum_{k=1}^{K} p(X \mid \omega_k)P(\omega_k)$

This requires that the set of classes are disjoint and complete. Ever sample belongs to one and only one class C_k.

Probability density functions are easily generalized to vectors of random variables.
Let $\vec{X} \in R^D$, be a vector random variables.
A probability density function, $p(\vec{X})$, is a function of a vector of continuous variables

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

The Central Limit theorem and Normal densities.

The "Central Limit Theorem" tells us that whenever the features an observation are the result of a sequence of N independent random events, the probability density of the features will tend toward a Normal or Gaussian density.

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. For any non-ideal density *p(X)* :

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

We can consider a sequence of random trials as a "source" of event

Source: S $\longrightarrow$ X

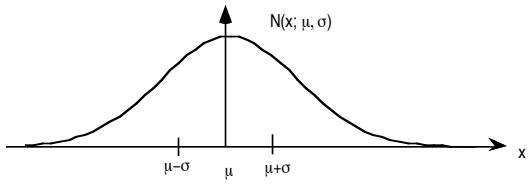
The central limit theorem tells us that in this case, a normalized sum of many independent random variables will converge to a Normal or Gaussian density function:

$$p(\vec{X}) = \mathcal{N}(\vec{X}; \vec{\mu}, \Sigma)$$

Univariate Normal Density Function

The Univariate (single variable) Gaussian density function is written:

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



The parameters of $\mathcal{N}(x;\ \mu,\ \sigma)$ are the first and second moments, μ and σ^2 of the function.

$$\mu = E\{X\} \qquad \sigma^2 = E\left\{(X-\mu)^2\right\} = E\left\{(X-E\{X\})^2\right\}$$

The normal function has 3 components:

1) Eulers Number "e" e = 2.718281828…

e is an irrational and transcendental constant approximately equal to 2.718281828….

Sometimes referred to as Euler's Number, e has many useful properties.
For use in the Normal density, e simplifies the algebra.

$$\int_{-\infty}^{\infty} e^x dx = e^x$$

2) a normalization factor:

$\frac{1}{\sqrt{2\pi}\sigma}$ is a normalization factor

$$\sqrt{2\pi}\sigma = \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

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

$$d(x,\mu;\sigma)^2 = \frac{(x-\mu)^2}{2\sigma^2}$$

This is the difference between x and μ normalized by the σ.

The normal density is e to the negative power of a distance.

To better understand, we need to talk about expected values, and moments.

The average value is the first moment of the samples

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.

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

The variance is the expected square of the deviation from the average

This is also the second moment of the pdf

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

Biased and Unbiased Variance

Note that this is a "Biased" variance. The unbiased variance would be

$$\tilde{\sigma}^2 = \frac{1}{M-1}\sum_{m=1}^{M}(X_m - \mu)^2$$

If we draw a random sample $\{X_m\}$ of M random variables from a Normal density with parameters (μ, σ)

$$\{X_m\} \leftarrow \mathcal{N}(x;\mu,\tilde{\sigma})$$

Then we compute the moments, we obtain.

$$\mu = E\{X_m\} = \frac{1}{M}\sum_{m=1}^{M} X_m$$

and

$$\hat{\sigma}^2 = \frac{1}{M}\sum_{m=1}^{M}(X_m - \mu)^2 \qquad \text{Where } \tilde{\sigma}^2 = \frac{M}{M-1}\hat{\sigma}^2$$

Note the notation: ~ means "true", ^ means estimated.

The expectation underestimates the variance by 1/M.

The RMS error for estimating *p(X)* from M samples $\{X_m\}$ is the difference between a biased and unbiased error. We can use the difference to estimate the sample error for using a biased estimate.

Multivariate Normal Density Function

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

Where the parameters $\vec{\mu}$, Σ and the mean and covariance of the density. These are the first and second moments of the density.

As above, we use upper case for probabilities and lower case for functions.
Thus $P(\omega)$ is a probability value, $p(X)$ is a function.

The mean is $\vec{\mu} = E\{\vec{X}\} = \begin{pmatrix} E\{X_1\} \\ E\{X_2\} \\ \dots \\ E\{X_D\} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_D \end{pmatrix}$

and the Covariance is $\Sigma = E\{(\vec{X} - E\{\vec{X}\})(\vec{X} - E\{\vec{X}\})^T\} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1D} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2D} \\ \dots & \dots & \ddots & \dots \\ \sigma_{D1} & \sigma_{D2} & \dots & \sigma_{DD} \end{pmatrix}$

where $\sigma_{ij} = \frac{1}{M_k}\sum_{m=1}^{M}(x_{mi} - \mu_{ki})(x_{mj} - \mu_{kj})$