Intelligent Systems: Reasoning and Recognition

James L. Crowley

Ensimag 2 Winter Semester 2020
Lesson 4 14 Feb 2020

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.

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

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

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

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

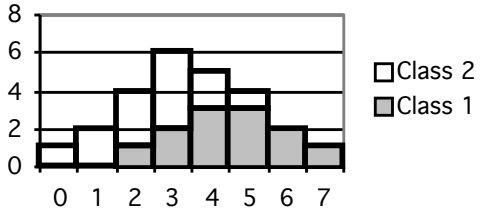
and $P(\omega_k)$ can be estimated from the relative size of the training set.

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

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).



Recall that the number of cells in the histogram is $Q=N^D$.
Having M >> Q is NECESSARY but NOT Sufficient.
Having M < Q is a guarantee of INSUFFICIENT TRAINING DATA.

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?

The worst case 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

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

The worst case occurs for a uniform probability density function.

For most applications, $M \geq 8Q$ (8 samples per "cell") is reasonable
(less than 12% RMS error).

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 $\vec{z}$.

The M training samples define M overlapping Parzen windows.
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.

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

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

$k(\vec{u}) \ge 0$ 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 $R = \arg\min_{\{X_s\}} \max\left\{\left\|\vec{X} - \vec{x}_s\right\|\right\}$

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

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

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: $p(\vec{X}) = \frac{S}{MV}$

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 $[-\infty, \infty]$
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

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

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 $[-\infty, \infty]$
2) $\int_{-\infty}^{\infty} p(\vec{x})d\vec{x} = 1$