

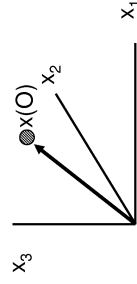
Computational Intelligence: Methods and Applications

Lecture 3
Histograms and probabilities.

Włodzisław Duch
SCE, NTU, Singapore
Google: Duch

Feature space representation

- Representation: mapping objects/states into vectors, $\{O\} \Rightarrow X(O)$, with $X_i(O)$ being i -th attribute of object O_i
- Attribute” and “feature” are used as synonyms, although strictly speaking “age” is an attribute, and “young” is its feature value.
- Types of features.
Categorical: symbolic or discrete – may be **nominal** (unordered), like “sweet, salty, sour”, or **ordinal** (can be ordered), like colors or small < medium < large (drink).
Continuous: numerical values.



Vector $X = (x_1, x_2, x_3 \dots x_d)$,
or a d -dimensional point
in the feature space.



Features

AI uses complex knowledge representation methods.
Pattern recognition is based mostly on simple feature spaces.

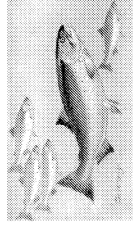
- Set of Objects: physical entities (images, patients, clients, molecules, cars, signal samples, software pieces), or states of physical entities (board states, patient states etc).
- Features: measurements or evaluation of some object properties.
- Ex: are pixel intensities good features?
No - not invariant to translation/scaling/rotation.



Better: type of connections, type of lines, number of lines ...
Selecting good features, transforming raw measurements that are collected, is very important.

Fishy example.

Chapter 1.2, Pattern Classification (2nd ed)
by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000



Singapore Fisheries automates the process of sorting salmon and sea bass fish, coming on a conveyor belt. Optical sensor and processing software evaluate a number of features: length, lightness, width, #fins

Step 1: look at the histograms.

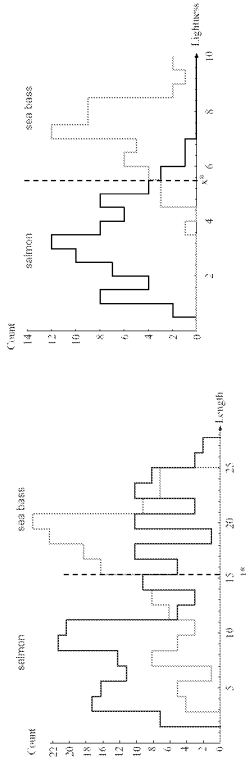
- Select number of bins, ex. $n=20$ (discretize data)
- calculate bin size $\Delta = (x_{max} - x_{min})/n$,
- calculate $N(C, r_i) = \#$ samples in class $C \in \{\text{salmon, bass}\}$ in each bin

$$r_i = [x_{min} + (i-1)\Delta, x_{min} + i\Delta], i=1 \dots n$$

- normalize $P(C, r_i) = N(C, r_i)/N$, where N is the number of all samples.
This gives joint probability $P(C, r_i) = P(r_i|C)P(C)$

Fishy histograms.

Example of histograms for two features, length (left) and skin lightness (right). Optimal thresholds are marked.



$P(r_i|C)$ is an approximation to the class probability distribution $P(x|C)$.

How to calculate it?

Discretization: replacing continuous values by discrete bins, or integration by summation, very useful when dealing with real data.

Alternative: integrate, fit some simple functions to these histograms, for example a sum of several Gaussians, optimize their width and centers.

Basic probability concepts.

Normalized co-occurrence (contingency) table: $P(C, r_i) = N(C, r_i) / N$

$$P(C, r_i) \begin{matrix} \text{matrix, columns=bins } r_i, \\ \text{rows = classes.} \end{matrix} \begin{pmatrix} P(C_1, r_1) & P(C_1, r_2) & P(C_1, r_3) \\ P(C_2, r_1) & P(C_2, r_2) & P(C_2, r_3) \\ P(C_3, r_1) & P(C_3, r_2) & P(C_3, r_3) \\ P(C_4, r_1) & P(C_4, r_2) & P(C_4, r_3) \\ P(C_5, r_1) & P(C_5, r_2) & P(C_5, r_3) \end{pmatrix}$$

$P(C, r_i)$ - joint probability distribution,

P of finding C and $x \in r_i$

$P(C)$ or a *priori* class probability, before making any measurements or learning that $x \in r_i$ is in some bin, is obtained by summing the row.

$$\sum_i P(C, x \in r_i) = P(C)$$

$P(x \in r_i)$ or probability that object from any class is found in bin r_i is obtained by summing the column.

$$\sum_j P(C_j, x \in r_i) = P(x \in r_i)$$

Fishy example.

Exploratory data analysis (EDA): visualize relations in the data.

How are histograms created?

Select number of bins, ex. $n=20$ (discretize data into 20 bins)

- calculate bin size $\Delta = (x_{max} - x_{min}) / n$,
- calculate $N(C, r_i) = \#$ samples from class C in each bin $r_i = [x_{min} + (i-1)\Delta, x_{min} + i\Delta], i=1 \dots n$

This may be converted to a joint probability $P(C, r_i)$ that a fish of the type C will be found within length in bin r_i

- $P(C, r_i) = N(C, r_i) / N$, where N is the number of all samples.

Histograms show joint probability $P(C, r_i)$ rescaled by N .

PDF and conditional probability

What if there x is a continuous variable and there are no natural bins?

Then $P(C, x)$ is a probability density function (pdf), and for small dx

$$P(C, [x, x+dx]) = P(C, x) dx$$

Suppose now that class C is known; what is the probability of finding $x \in r_i$ or for continuous features finding it in $[x, x+dx]$ interval?

$P(x \in r_i | C)$ denotes conditional probability distribution, knowing C .

Because sum over all bins gives: $\sum_i P(x \in r_i | C) = 1$

and for joint probability $\sum_i P(C, x \in r_i) = P(C)$ therefore the formula is

$$P(x \in r_i | C) = P(C, x \in r_i) / P(C)$$

$P_C(x) = P(x|C)$ class probability distribution is simple rescaled joint probability, divide a single row of $P(C, x)$ matrix by $P(C)$.

Probability formulas

Most probability formulas are simple summations rules!

Matrix of joint probability distributions:

elements in $P(C, x)$ for discrete x , or $P(C, x \in r_i)$ after discretization.

Just count how many times $N(C, x)$ is found.

$$P(C, x) = N(C, x)/N$$

$$P(C) = \sum_{i=1}^n P(C, x_i);$$

Row of $P(C, x)$ sums to:

$$\sum_{i=1}^n P(x_i | C) = 1;$$

therefore $P(x|C) = P(C, x)/P(C)$
sums to

For x continuous $P(C, x)$ is the probability

density distribution, integrating to $P(C)$

$$P(C) = \int P(C, x) dx$$

Bayes formula

Bayes formula allows for calculation of the conditional probability distribution $P(x|C)$ and posterior $P(C|x)$ probability distribution.

These probabilities are renormalized elements of the joint probability $P(C, r_i)$.

$$\sum_C P(C) = 1; \quad \int P(x) dx = 1$$

$$\sum_i P(x_i) = 1$$

They sum to 1 since we know that the object with $x \in r_i$ is from C class, and that given x it must be in one of the C classes.

$$\sum_i P(x_i | C) = \sum_C P(C | x) = 1;$$

$$P(x_i | C) = P(C, x_i) / P(C)$$

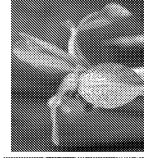
$$P(C | x_i) = P(C, x_i) / P(x_i)$$

$$P(C | x_i) P(x_i) = P(x_i | C) P(C)$$

Therefore Bayes formula is quite obvious!

Example

Two types of Iris flowers:
Iris Setosa and Iris Virginia



Measuring petal lengths in cm in two intervals, $r_1 = [0, 3]$ cm and $r_2 = [3, 6]$ cm of 100 samples we may get for example the following distribution:

$$N(C, r) = \begin{pmatrix} 36 & 4 \\ 8 & 52 \end{pmatrix} \quad \begin{aligned} N(C_1) &= 40, N(C_2) = 60 \\ N(r_1) &= 44, N(r_2) = 56 \end{aligned}$$

Therefore probabilities for finding different types of Iris flowers is:

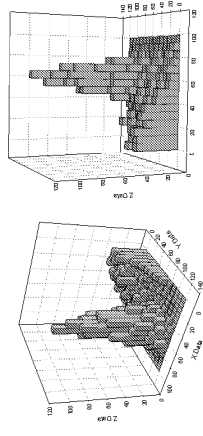
$$P(C, r) = \begin{pmatrix} 0.36 & 0.04 \\ 0.08 & 0.52 \end{pmatrix} \quad \begin{aligned} P(C_1) &= 0.4; P(r_1) = 0.44 \\ P(C_2) &= 0.6; P(r_2) = 0.56 \end{aligned}$$

Calculate conditional probabilities and verify all formulas given on the previous pages!

Do more examples.

Going to 2D histograms

2D histograms in 3D: still useful, although sometimes hard to analyze. For a single class it is still easy, but for >2 classes rather difficult.



Many visualization software packages create such 3D plots.

Joint probability $P(C, x, y)$ is shown here, for each class on separate drawing; for small N it may be quite different than real distribution.

