Computational Intelligence: Methods and Applications

Lecture 26

Density estimation,
Expectation Maximization.

Włodzisław Duch
SCE, NTU, Singapore
Google: Duch

Density estimation

Knowledge of joint probability density $P(C,X)$ or just $P(X)$ allows to do much more than just discrimination!

Local maxima of probability density functions (PDFs) correspond to combination of features defining objects in feature spaces.

Estimating PDFs we may create adaptive systems learning from data with or without supervision. They are useful for:

- Auto-association and hetero-association.
- Completion of unknown parts of the input vector (content-addressable memory), prediction of missing values.
- Extraction of logical rules, classical and probabilistic (or fuzzy).
- Finding prototypes for objects or categories in feature spaces.
- Using density functions as heuristics for solution of complex problems, learning from partial info & solving complex problems.

Cognitive inspirations

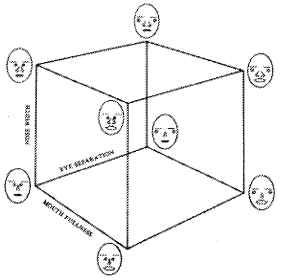
How do we recognize objects? Nobody really knows ...

Objects have features, combinations of features, or rather distributions of feature values in Feature Spaces (FS), characterize objects.

A single object is a point in the FS;

similar objects create a category, or a concept: for ex. happy or sad face, corresponding to some area of the feature space.

P(Angry|Face features) will have maximum around one of the corners.

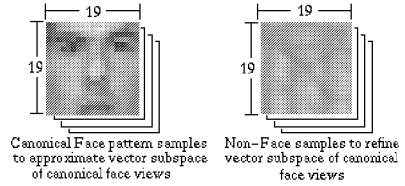


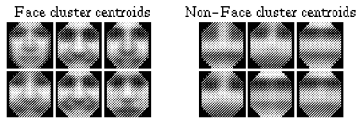
In cognitive psychology FS are called “psychological spaces”.

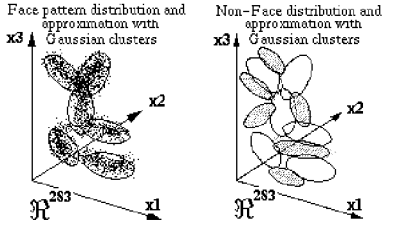
The shape of the $P(X|C)$ distribution may be quite complex, estimated using known samples to create a fuzzy prototype.

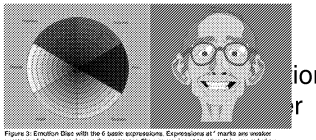
Object recognition

Population of neural columns, each acting as a weak classifiers to recognize some features, working in chorus – similar to “stacking”. Second-order similarity in low-dimensional (<300) space is sufficient.









ion
r

processes are much more complex, results of neurodynamics may be approximated by PDFs.

Missing features

Suppose that one of the features X =(X_1,X_2, ... X_d), for example X_1, is missing. What is the most likely value for this feature?

Frequently an average value $E(X_1)$ is used, but is this a reasonable idea? The average may fall in an area where there is no data!

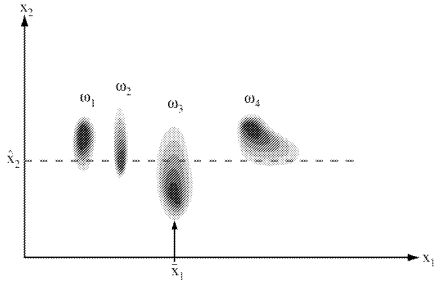


Fig. 2.22, Duda, Hart & Stork, 2000

In this case if X_2 is known the best answer is the value corresponding to the maximum density at ω_2.
Recover missing values searching for maximum density!

Maximum likelihood

Suppose that the density $P(\mathrm{X};\theta)$ is approximated using a combination of some parameterized functions. Given a set of observations (data samples) $D=\{\mathbf{X}^{(i)}\}$, $i=1..n$, what parameters should one choose?

Parameters θ may include also missing values, as a part of the model.

A reasonable assumption is that the observed data *D* should have high chance of being generated using the model $P(D;\theta)$. Assuming that the data vectors $\mathrm{X}^{(i)}$ are independent, the likelihood of obtaining the dataset *D* is:

$$l(\theta;D)=\prod_{i=1}^{n} P\left(\mathbf{X}^{(i)};\theta\right)$$

The most probable parameters of the model (including missing values) maximize likelihood. To avoid products use logarithm and minimize -*L*

$$\min_{\theta} L(\theta;D)=-\min_{\theta}\ln l(\theta;D)=-\sum_{i=1}^{n}\ln P\left(\mathbf{X}^{(i)};\theta\right)$$

Solution

Maximum is found by setting the derivative of the log-likelihood to 0:

$$\frac{\partial L(\theta)}{\partial\theta}=-\sum_{i=1}^{n}\frac{1}{P\left(\mathbf{X}^{(i)};\theta\right)}\frac{\partial P\left(\mathbf{X}^{(i)};\theta\right)}{\partial\theta}=0$$

Depending on the parameterization, sometimes this can be solved analytically, but for almost all interesting functions (including Gaussians) iterative numerical minimization methods are used.

Many local minima of the likelihood function are expected, so the minimization problem may be difficult.

Likelihood estimation may be carried for samples from a given class $P(\mathrm{X}|\omega;\theta)$, assuming that the probability of generating *n* such samples is equal to $P(\mathrm{X}|\omega;\theta)^n$, and the *a priori* class probabilities are estimated from their frequencies.

Such parametric models are called “generative” models.

Example

Example from “Maximum likelihood from incomplete data via the EM algorithm”, Dempster, Laird, Rubin 1977, data by Rao, from population genetics. There are 197 observation of 4 types of bugs:

n_1=125 times species (class) ω_1, n_2 = 18 from class ω_2, n_3 =20 from class ω_3, and n_4 =34 from class ω_4. An expert provided the following parametric expressions for the probabilities to find these bugs:

$$P\left(\mathbf{X}\mid\omega_1;\theta\right)=\left(2+\theta\right)/4;\ P\left(\mathbf{X}\mid\omega_2;\theta\right)=\left(1-\theta\right)/4$$
$$P\left(\mathbf{X}\mid\omega_3;\theta\right)=\left(1-\theta\right)/4;\ P\left(\mathbf{X}\mid\omega_4;\theta\right)=\theta/4$$

Find the value of parameter that maximize the likelihood:

$$l(\theta)=P\left(\mathbf{X}\mid\omega_1;\theta\right)^{n_1}P\left(\mathbf{X}\mid\omega_2;\theta\right)^{n_2}P\left(\mathbf{X}\mid\omega_3;\theta\right)^{n_3}P\left(\mathbf{X}\mid\omega_4;\theta\right)^{n_4}$$

Multiplicative constants $n!/(n_1!n_2!n_3!n_4!)$ are not important here.

Solution

Log–likelihood: $$L(\theta) = -n_1 \ln P(\mathbf{X} \mid \omega_1; \theta) - n_2 \ln P(\mathbf{X} \mid \omega_2; \theta) - n_3 \ln P(\mathbf{X} \mid \omega_3; \theta) - n_4 \ln P(\mathbf{X} \mid \omega_4; \theta)$$

Derivative: $$\frac{\partial L(\theta)}{\partial \theta} = -\left(\frac{n_1}{2+\theta} - \frac{n_2 + n_3}{1-\theta} + \frac{n_4}{\theta} \right) = 0$$

Quadratic equation for θ allows for analytical solution: θ = 0.6268; now the model may provide estimations of expected frequencies:

$$\langle n_1 \rangle = N(2+\theta)/4 = 129.4$$

For all 4 classes, expected (real) number of observation:

<n_1>=129 (125), <n_2>=18 (18), <n_3>=18 (20), <n_4>=31 (34)

In practice analytic solutions are rarely possible.

General formulation

Given data vectors $D=\{\mathbf{X}^{(i)}\}$, $i=1..n$, and some parametric functions $P(\mathrm{X}|\theta)$ that model the density of the data $P(\mathrm{X})$ the best parameters should minimize log-likelihood for all data samples:

$$\theta^* = \arg\min_\theta L(\theta \mid D) = -\sum_{i=1}^{n} \ln P(\mathbf{X}^{(i)}; \theta)$$

$P(\mathrm{X}|\theta)$ is frequently a Gaussian mixture; for a single Gaussian standard solution will give the formula for mean and variance.

Assume now that X is not complete – features, or whole parts of the vector are missing. Let Z=(X,Y) be the complete vector. Joint density:

$$P(\mathbf{Z} \mid \theta) = P(\mathbf{X}, \mathbf{Y} \mid \theta) = P(\mathbf{Y} \mid \mathbf{X}, \theta) P(\mathbf{X} \mid \theta)$$

Initial joint density may be formed analyzing cases without missing values; the idea is to maximize the complete data likelihood.

What to expect? E-step.

Original likelihood function $L(\theta|\mathbf{X})$ is based on incomplete information, and since Y is unknown it may be treated as a random variable that should be estimated.

Complete-data likelihood function $L(\theta|\mathbf{Z})=L(\theta|\mathbf{X},\mathbf{Y})$ may be evaluated calculating the expectation of incomplete likelihood over Y. This is done iteratively, starting from initial estimation θ^{i-1} new estimation θ^{i} of parameters and missing values is generated:

$$Q(\theta \mid \theta^{i-1}) = E_{\mathbf{Y}}\left[\ln P(\mathbf{X}, \mathbf{Y} \mid \theta) \mid \mathbf{X}, \theta^{i-1} \right]$$

where X and θ^{i-1} are fixed, θ is a free variable, and the conditional expectation is calculated using the joint distribution of the X, Y variable with fixed X

$$E[Y \mid X = x] = \int y P_{Y|X}(x, y) dy$$

See detailed ML discussion in Duda, Hart & Stork, Chapter 3