Theory of marginal mixture estimation

Marginal mixture estimation deals with explanatory variables $x = [x_1, x_2, \cdots x_n]$ affecting the target (modeled) variable y. It performs an analysis of the space of explanatory variables x to build local models describing the dependence of y on x. To this end, we consider the full model of data in the form of joint distribution with its factorization according to the chain rule

$$f(y, x) = f(x) f(y|x).$$

The first distribution on the right hand side describes the space x which is analyzed by clustering. Based on this analysis, the model describing the influence of x on y is constructed. It is represented by the second distribution

The main assumption is that the variables in x can be modeled as independent mixtures interconnected only by their pointers.

The method has three parts

1. **Mixture estimation** of the model $f(x)$ creating clusters in the space x. These clusters are covered by components $f_{i,j}(x)$ where i denotes the variable x_i and j denotes the j-th component. This part runs on-line.

2. **Construction of local models** in each variable x_i and each cluster $C_{i,j}$ of this variable. The local model is based only on the data from the cluster, i.e. on the data which are classified to the corresponding component. This part can be accomplished off-line.

3. **Classification** of a newly coming data record x_t into the classes given by the values of y. This is performed as zero-step prediction of y_t corresponding to the measured data x_t. The value $\hat{y}_t$ of the point estimate y_t determines the class into which the x_t is classified.

Now we show the background of individual steps in comparison with the standard way of mixture estimation.

Part 1

Clustering in the space of data x

Standard mixture estimation

Let us recall. The weights for statistics update are proportional to the component with inserted measured data x_t and the actual value of the point estimate $\hat{\theta}_{j;t-1}$ of component parameter θ_j.

$$w \propto q_j = f\left(x_t|\hat{\theta}_{j;t-1}\right) \tag{1}$$

The update of statistics (e.g. for $S = \mathrm{sum}\,(y)$ and $\kappa = \mathrm{count}\,(y)$) is

$$\begin{aligned} S_{j;t} &= S_{j;t-1} + w_j y_t \\ \kappa_{j;t} &= \kappa_{j;t-1} + w_j \end{aligned} \tag{2}$$

From the updated statistics we construct the updated values of parameters.

Estimation of marginal mixtures

It copies the standard way with some differences. And these are specific for both the suggested methods: common and different components.

The common feature is that (for multivariate x) we work with local components

$$f_j\left(x_i|\theta_{ij}\right)$$

where j denotes the component and i variable the x_i.

The main difference between the introduced methods (common and different) is that with common components, there exist co called overall components composed of the local ones

$$f_j\left(x|\theta_j\right) = \prod_{i=1}^{n_c} f_j\left(x_i|\theta_{ij}\right)$$

on condition of independency of variables x_i. In the method with different components, the overall component can be composed from arbitrary combination of local components from individual variables.

From it follows that with common method, the number of components in each variable must be the same. The different method can have different numbers of local components in each variable.

The differences between the methods are shown in the following table:

Common components	**Different components**

Proximities in both cases are computed from the local models, for each variable and each its component

$$q_{ij} = f_j\left(x_i|\hat{\theta}_{ij;t-1}\right)$$

The overall component weight are $$W_j \propto \prod_{i=1}^{n} q_{ij}$$	We work with the local proximities $$w_{ij} \propto \frac{q_{ij}}{\sum_k q_{ik}}$$ for each i normalized over j.
Update of the statistics (for S and κ) $$S_{ij;t} = S_{ij;t-1} + W_j y_t$$ $$\kappa_{ij;t} = \kappa_{ij;t-1} + W_j$$	Update of the statistics (for S and κ) $$S_{ij;t} = S_{ij;t-1} + w_{ij} y_t$$ $$\kappa_{ij;t} = \kappa_{ij;t-1} + w_{ij}$$

Graphical illustration of both methods is here

Part 2

Local explanatory models

Construction of local explanatory models $f_j\left(x_i|\vartheta_{ij}\right)$ to be used for prediction $f\left(y|x\right)$ via Naive Bayes methodology.

Each local component defines its data cluster in the respective variable. It is formed by records x_i for which the component was active (had the greatest weight) and corresponding values of the target variable y . On the data from these clusters we estimate off-line (as a normalized frequency table) the categorical local models.

From the variable y we determine the probability function $f\left(y\right)$, simply as a normalized histogram.

Part 3

Classification

Here, we practically repeat the procedure from estimation (Part 1) with the only difference that we do not update the component statistics and use the models estimated in Part 1. I.e. for the measured data record $x = [x_1, x_2, \cdots x_i, \cdots x_n]$ we determine the actual weights w_{ij} using the local models $f_j\left(x_i\right)$ from Part 1 with the fixed parameters. Then, using the constructed weights and local models from Part 2, we construct the predictive probability function $f\left(y|x_t\right)$ for measured value x_t. Again, the methods (common and different) a bit differ.

Common components	**Different components**
We measure a new value x_t (which will be classified into the class given by the value of y)	
The overall component weight are $$W_j \propto \prod_{i=1}^{n} q_{ij}$$	We work with the local proximities $$w_{ij} \propto \frac{q_{ij}}{\sum_k q_{ik}}$$ for each i normalized over j.
Predictive pf construction $$f_j\left(y|x_t\right) \propto \prod_i f_j\left(x_{i;t}|y\right) f\left(y\right)$$ $$f\left(y|x_t\right) = \sum_{j=1}^{n_c} f_j\left(y|x_t\right)$$ i.e. first we construct overall components over all variables applying Naive Bayes and then we combine mixture .	Predictive pf construction $$f\left(x_{i;t}|y\right) = \sum_{j=1}^{n_c(i)} w_{ij} f_j\left(x_i|y\right)$$ $$f\left(y|x_t\right) \propto \prod_i f\left(x_{i;t}|y\right) f\left(y\right)$$ i.e. first we construct the mixture of components within each variable and then we use Naive Bayes over variables.

Program to estimation of mixtures with common components

Program and its description

Program to estimation of mixtures with different components

Program and its description