Bayesian Classifiers with Continuous Predictors

Pedro Larrañaga, Concha Bielza, Jose Luis Moreno

Computational Intelligence Group
Artificial Intelligence Department
Universidad Politécnica de Madrid

Machine Learning

Master in Data Science + Master HMDA

Outline

Outline

Gaussian naive Bayes classifier (Friedman et al., 1998)

Gaussian naive Bayes classifier

- The conditional density of each predictor variable X_i, given a value of the class variable, c, follows a Gaussian distribution:
  $X_i|C = c \sim \mathcal{N}(x_i|\mu_{c,i}, \sigma_{c,i})$ for all $i = 1, \ldots, n$; $c = 1, ..., R$

$$c^* = \arg\max_c p(c) \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi}\sigma_{c,i}} e^{\frac{-1}{2}\left(\frac{x_i - \mu_{c,i}}{\sigma_{c,i}}\right)^2} \right]$$

- The total number of parameters to be estimated is: $(R-1) + 2nR$. Maximum likelihood for estimations
- This model is equivalent to a particular case of a quadratic discriminant analysis with a diagonal covariance matrix Σ_r for each class c_r

Gaussian naive Bayes classifier

Filter-wrapper selective Gaussian naive Bayes classifier

$$\mathbb{I}(X_i, C) = \frac{1}{2}[\log_2(\sigma_i^2 - \sum_{c=1}^{R} p(c)\log_2(\sigma_{c,i})^2)]$$

Mutual information between a Gaussian variable and a categorical one

1. $\mathbb{I}(X_i, C)$ $(i = 1..., n)$ is used to sort the predictor variables in descending order The set of predictor variables is initialized to the empty set, and all instances are classified in the most frequent class
2. Variables are added one by one in order of mutual information to the set of predictor variables, and the accuracy of the model is estimated
3. Outputs the selective classifier associated with the variable subset that has achieved the best estimated accuracy in the search process

The upper bound for the number of models that we can build is n

Filter selective Gaussian naive Bayes classifier

Induces the classifier with the subset of variables $\{X_{(1)}, \ldots, X_{(h)}\}$, where h is the last order for which $\mathbb{I}(X, C) > t$, and t denotes the a priori fixed threshold for the selection of variables

Gaussian semi-naive Bayes classifier

Gaussian semi-naive Bayes classifier

- The conditional density for each joint variable $\mathbf{Y}$ with m components is given by $f(\mathbf{y}|c) = (2\pi)^{-\frac{1}{2}m}|\mathbf{\Sigma}_c|e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_c)^T(\mathbf{\Sigma}_c)^{-1}(\mathbf{y}-\boldsymbol{\mu}_c)}$, where $\mathbf{\Sigma}_c$ and $\boldsymbol{\mu}_c$ are the covariance matrix and mean vector of $\mathbf{Y}$ conditioned on a class value c
- The forward sequential selection and joining is applicable

Wrapper condensed backward semi-naive Bayes (Perez et al., 2006)

- Wrapper greedy backward algorithm using a selection of the predictor variables as a multidimensional joint Gaussian at each step
- At the beginning, all predictor variables are in the model
- At each step of the algorithm, one variable is chosen for exclusion
- This process is repeated until further removals fail to improve accuracy

Gaussian tree-augmented naive Bayes classifier

Gaussian tree-augmented naive Bayes classifier. Filter

- Adaptation of the TAN classifier to the situation where X_i and X_j, conditioned on each value c of variable C, follow a bivariate normal density
- The conditional mutual information between X_i and X_j given C is computed as

$$\mathbb{I}(X_i, X_j|C) = -\frac{1}{2}\sum_{c=1}^{R} p(c)\log_2(1-\rho_c^2(X_i, X_j))$$

where $\rho_c^2(X_i, X_j)$ denotes the correlation coefficient of X_i and X_j when $C = c$

Outline

Kernel density estimation (Silverman, 1986)

Kernel density estimation (Silverman, 1986)

- The general form of a kernel-based n-dimensional estimator is

$$f_{\text{Kernel}}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \kappa_{\mathbf{H}}(\mathbf{x} - \mathbf{x}^i)$$

  where $\mathbf{H}$ is an $n \times n$ bandwidth matrix and $\kappa_{\mathbf{H}}$ is the kernel function
- The kernel-based density estimator $f_{\text{Kernel}}(\mathbf{x})$ is determined by averaging N kernel functions $\kappa_{\mathbf{H}}(\mathbf{x} - \mathbf{x}^i)$ placed at each observation $\mathbf{x}^i$
- The kernel function $\kappa_{\mathbf{H}}$ is defined as $\kappa_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{\frac{1}{2}} K(\mathbf{H}^{\frac{1}{2}}\mathbf{x})$ where K is an n-dimensional density function
- The kernel density estimate is built by centering a scaled kernel at each instance of the data set, and can be seen as a sum of bumps placed at these instances
- The value of the kernel estimate at point $\mathbf{x}$ is computed as the average of the N kernels at that point
- An example of a kernel function is an n-dimensional Gaussian density centered at $(0, 0, ..., 0)$ with the identity as variance-covariance matrix, that is, $K(\mathbf{x}) = (2\pi)^{-n/2} \exp(-\frac{1}{2}\mathbf{x}^T\mathbf{x})$

Kernel density estimation (Silverman, 1986)

The bandwidth matrix H

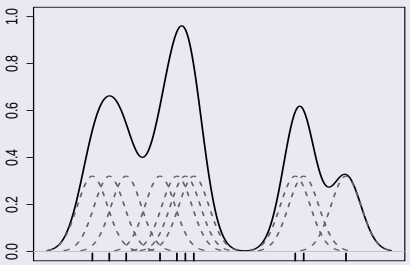
- The kernel function $\kappa_{\mathbf{H}}(\mathbf{x})$ determines the shape of the bumps
- The bandwidth matrix **H** establishes the degree of smoothing of the kernel-based n-dimensional estimator
- A good selection of **H** is crucial
- For n dimensional kernel-based estimators, the number of parameters required to specify a full bandwidth matrix is quadratic in n
- A simple way to estimate **H** is using the differential scaled approach (Simonoff, 1996) which depends on a unique smoothing parameter h
- The differential scaled method considers **H** as a diagonal matrix, whose j-th element is computed as $h^2 s_j^2$, with s_j^2 being the sample standard deviation of X_j
- The normal rule is often used for determining the h value

$$h = \left(\frac{4}{(m+2)N} \right)^{\frac{1}{m+4}}$$

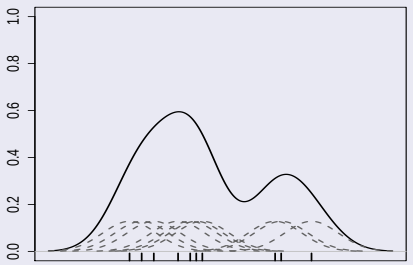
where m is the number of continuous variables to be estimated by the kernel density

Kernel density estimation (Silverman, 1986)

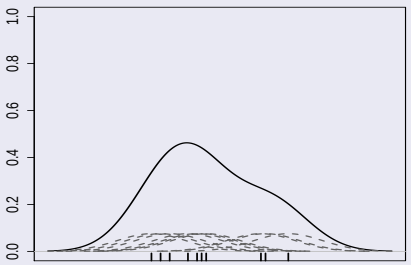
Effects in the kernel density estimator of the smoothing degree controlled by parameter h



(a) h under the optimum



(b) Close to optimum value of h



(c) h value over the optimum

Kernel-based classifiers

Kernel-based classifiers as flexible Bayesian classifiers

- To transform standard Bayesian classifiers into kernel-based Bayesian classifiers, it is enough to be able to compute the conditional (given the class variable) mutual information of the pairs of continuous variables (**X** and **Y**) of any dimensionality

$$\mathbb{I}(\mathbf{X},\mathbf{Y}|C) = \sum_{c=1}^{R} p(c) \sum_{i=1}^{N_c} \log_2 \frac{f_{\text{Kernel}}(\mathbf{x}^i,\mathbf{y}^i|C=c)}{f_{\text{Kernel}}(\mathbf{x}^i|C=c) f_{\text{Kernel}}(\mathbf{y}^i|C=c)}$$

  N_c is the number of instances satisfying $C = c$ and superindex i refers to the i-th instance in the partition induced by the value c of the class variable

- Flexible naive Bayes (John and Langley, 1995)
- Flexible tree-augmented naive Bayes (Perez et al., 2009)
- The Parzen window classifier (Parzen, 1962) as a kernel-based Bayesian classifier with a complete graph representing the relationships between continuous variables

Outline

Bayesian classifiers with mixed predictors

Naive Bayes classifier with discrete and continuous predictors

For the case of categorical and univariate Gaussians, the class value c^* is computed as the maximum a posteriori

$$c^* = \arg\max_c p(c|\mathbf{x},\mathbf{y}) = \arg\max_c p(c) \prod_{i=1}^{n_1} p(x_i|c) \prod_{j=1}^{n_2} \left[\frac{1}{\sqrt{2\pi}\sigma_{c,j}} e^{-\frac{1}{2}\left(\frac{y_j-\mu_{c,j}}{\sigma_{c,j}}\right)^2} \right]$$

where $\{X_1, ..., X_{n_1}\}$ and $\{Y_1, ..., Y_{n_2}\}$ denote the subsets of discrete and continuous predictor variables, respectively

References

- Friedman N, Goldszmidt M, Lee T (1998). Bayesian network classification with continuous attibutes: Getting the best of both discretization and parametric fitting. *Proceedings of the 15th National Conference on Machine Learning*, 179-187
- John G H, Langley P (1995). Estimating continuous distributions in Bayesian classifiers. *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence*, 338-345,
- Parzen E (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*,33(3), 1065-1076
- Pérez A, Larrañaga P, Inza I (2006). Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes. *International Journal of Approximate Reasoning*, 43, 1-25
- Silverman B W (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall
- Simonoff J S (1996). *Smoothing Methods in Statistics*. Springer
- Pérez A, Larrañaga P, Inza I (2009). Bayesian classifiers based on kernel density estimation: Flexible classifiers. International Journal of Approximate Reasoning, 50, 341-362

BAYESIAN CLASSIFIERS WITH CONTINUOUS PREDICTORS

Pedro Larrañaga, Concha Bielza, Jose Luis Moreno

Computational Intelligence Group
Artificial Intelligence Department
Universidad Politécnica de Madrid

Computational Intelligence Group

Departamento Inteligencia Artificial

Machine Learning

Master in Data Science + Master HMDA