Naive Bayes

Andrea Passerini
passerini@disi.unitn.it

Machine Learning

Naive Bayes classifier

Setting

- Each instance x is described by a conjunction of attribute values $\langle a_1, \dots, a_m \rangle$
- The target function can take any value from a finite set of $\mathcal{Y}$
- The task is predicting the MAP target value given the instance:

$$y^* = \text{argmax}_{y_i \in \mathcal{Y}} P(y_i|x) = \text{argmax}_{y_i \in \mathcal{Y}} \frac{P(a_1, \dots, a_m|y_i)P(y_i)}{P(a_1, \dots, a_m)}$$

$$= \text{argmax}_{y_i \in \mathcal{Y}} P(a_1, \dots, a_m|y_i)P(y_i)$$

Naive Bayes assumption

Learning problem

Class conditional probabilities $P(a_1, \ldots, a_m | y_i)$ are hard to learn, as the number of terms is equal to the number of possible instances times the number of target values

Simplifying assumption

- Attribute values are assumed independent of each other given the target value:

$$P(a_1, \ldots, a_m | y_i) = \prod_{j=1}^{m} P(a_j | y_i)$$

- Parameters to be learned reduce to the number of possible attribute values times the number of possible target values

Naive Bayes classifier

definition

$$y^* = \mathrm{argmax}_{y_i \in \mathcal{Y}} \prod_{j=1}^{m} P(a_j|y_i)P(y_i)$$

Single distribution case

- Assume all attribute values come from the same distribution.
- The probability of an attribute value given the class can be modeled as a multinomial distribution over the K possible values:

$$P(a_j|y_i) = \prod_{k=1}^{K} \theta_{ky_i}^{z_k(a_j)}$$

Naive Bayes classifier

Parameters learning

- Target priors $P(y_i)$ can be learned as the fraction of training set instances having each target value
- The maximum-likelihood estimate for the parameter θ_{kc} (probability of value v_k given class c) is the fraction of times the value was observed in training examples of class c:

$$\theta_{kc} = \frac{N_{kc}}{N_c}$$

- Assume a Dirichlet prior distribution (with parameters $\alpha_{1c}, \ldots, \alpha_{Kc}$) for attribute parameters.
- The posterior distribution for attribute parameters is again multinomial:

$$\theta_{kc} = \frac{N_{kc} + \alpha_{kc}}{N_c + \alpha_c}$$

Example: text classification

Task

- Classify documents in one of C possible classes.
- Each document is represented as the *bag-of-words* it contains (i.e. no position information)
- Let V be the vocabulary of all possible words
- A dataset of labeled documents $\mathcal{D}$ is avaiable

Example: text classification

Naive Bayes learning

- Compute prior probabilities of classes as: $P(y_i) = \frac{|\mathcal{D}_i|}{|\mathcal{D}|}$ where $\mathcal{D}_i$ is the subset of training examples with class y_i.
- Model attributes with a multinomial distribution with $K = |V|$ possible states (words).
- Compute probability of word w_k given class c as the fraction of times the word appear in documents of class y_i, wrt to all words in documents of class c:

$$\theta_{kc} = \frac{\sum_{\boldsymbol{x} \in \mathcal{D}_c} \sum_{j=1}^{|\boldsymbol{x}|} z_k(x[j])}{\sum_{\boldsymbol{x} \in \mathcal{D}_c} |\boldsymbol{x}|}$$

Example: text classification

Naive Bayes classification

$$
\begin{aligned}
y^* &= \mathrm{argmax}_{y_i \in \mathcal{Y}} \prod_{j=1}^{|\boldsymbol{x}|} P(x[j] | y_i) P(y_i) \\
&= \mathrm{argmax}_{y_i \in \mathcal{Y}} \prod_{j=1}^{|\boldsymbol{x}|} \prod_{k=1}^{K} \theta_{k y_i}^{z_k(x[j])} \frac{|\mathcal{D}_i|}{|\mathcal{D}|}
\end{aligned}
$$

Naive Bayes classifier

Note

- We are making the simplifying assumption that all attribute values come from the same distribution
- Otherwise attributes from different distributions have to be considered separately for parameter estimation

Example

- Assume each instance x is represented as a vector of ℓ attributes
- Assume the j^{th} attribute ($j \in [1, \ell]$) can take $\{v_{j1}, \ldots, v_{jK_j}\}$ possible values.
- The parameter θ_{jkc} representing the probability of observing value v_{jk} for the j^{th} attribute given class c is estimated as:

$$\theta_{jkc} = \frac{\sum_{\boldsymbol{x} \in \mathcal{D}_c} z_{jk}(x[j])}{|\mathcal{D}_c|}$$