Machine Learning I WS2016/17

Lecturer: Prof. Dr. Bethge

3. (Gaussian) Naive Bayes

Task 1: Basic concepts

Naive Bayes reduces the number of parameters that must be estimated for a Bayesian classier, by making a conditional independence assumption when modeling $P(X|Y)$. The definition for conditional independence is the following:

Definition: Given random variables X, Y and Z, X is conditionally independent of Y given Z, denoted by $X \perp Y|Z$, if and only if:

$$P(X = x_i|Y = y_j, Z = z_k) = P(X = x_i|Z = z_k) \quad \forall i, j, k \tag{1}$$

Given this definition, please answer the following questions:

a. Given $X \perp Y|Z$, can we say $P(X, Y|Z) = P(X|Z)P(Y|Z)$? Explain.

b. Given $X \perp Y|Z$, can we say $P(X, Y) = P(X)P(Y)$? Explain.

c. Suppose X is a vector of n boolean attributes and Y is a single discrete-valued variable that can take on J possible values. Let $\theta_{ij} = P(X_i|Y = y_j)$. What is the number of independent θ_{ij} parameters?

d. Consider the same problem, but now suppose X is a vector of n real-valued attributes, where each of these X_i follows a Normal (Gaussian) distribution: $P(X_i = x_i|Y = y_j) \sim \mathcal{N}(x_i|\mu_{ij}, \sigma_{ij})$. How many distinct μ_{ij}, σ_{ij} are there?

We can write the classification rule for Naive Bayes as:

$$y^* = \text{argmax}_{y_k} \frac{P(Y = y_k) \prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i|Y = y_j)}. \tag{2}$$

e. We often do not compute the denominator when estimating Y. Explain why.

f. Is it possible to calculate $P(X)$ from the parameters estimated by Naive Bayes?

Task 2: Parameter estimation for Naive Bayes

Whether X takes discrete or continuous inputs, Naive Bayes can be used for classication with the same conditional independence assumptions. In this question, we'll discuss how to estimate the parameters using MLE for both of the cases.

1. Let $X = \langle X_1, X_2, \ldots, X_n \rangle$ be a vector of n Boolean values where the random variable X_i denotes the ith attribute of X. Suppose we are interested in estimating the parameters for the first attribute X_1. We typically model $P(X_1|Y = y_k)$ with a Bernoulli distribution:

$$P(X_1 = x_{1j}|Y = y_k) = \theta_{1k}^{x_{1j}}(1 - \theta_{1k})^{(1-x_{1j})} \tag{3}$$

where $j = 1, \ldots, M$ refers to the jth training instance (M is the number of training samples), and where x_{1j} refers to the value of X_1 in the jth training instance. Assume that the M training instances are independent and identically distributed (iid). Derive the MLE for θ_{1k}.

2. Now suppose each X_i is distributed normally, i.e

$$P(X_i = x_{ij}|Y = y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} e^{-\left(\frac{x_{ij}-\mu_{ik}}{2\sigma_{ik}}\right)^2}. \tag{4}$$

Suppose the variance is independent of the class variable Y, and X_i, i.e. $\sigma_{ik} = \sigma$. Derive the MLE estimator for μ_{ik}.

Task 3: Conjugate prior of multinomial distribution

Show that the Dirichlet distribution

$$P(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{m} \theta_k^{\alpha_k - 1} \tag{5}$$

is the conjugate prior of the multinomial distribution

$$P(x|\theta) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{k=1}^{m} \theta_k^{x_k}. \tag{6}$$

Task 4: Apply Naive Bayes on Reuters dataset

We will now learn how to use Naive Bayes to solve a real world problem: text categorization. Text categorization (also referred as text classification) is the task of assigning documents to one or more topics. For our homework, we will use a benchmark dataset that is frequently used in text categorization problems. This dataset, Reuters-21578, consists of documents that were appeared in Reuters newswire in 1987. Each document was then manually categorized into a topic among over 100 topics. In this homework we are only interested in earn and acquisition (acq) topics, so we will be using a shortened version of the dataset (documents assigned to topics other than "earn" or "acq" are not in the dataset provided for the homework). As features, we will use the frequency (counts) of each word occurred in the document. This model is known as bag of words model and it is frequently used in text categorization. You can download the HW2 data from *goo.gl/iHkbfw*. In this folder you will find:

train.csv: Training data. Each row represents a document, each column separated by commas represents features (word counts). There are 4527 documents and 5180 words.
train labels.txt: labels for the training data
test.csv: Test data, 1806 documents and 5180 words
test labels.txt: labels for the test data
word indices: words corresponding to the feature indices.

For your convenience we also included a version of this dataset in .mat format, (reuters.mat) so that you can directly import it to Matlab and Python. Implement Naive Bayes. To avoid 0 probabilities, choose a Beta distribution with equal valued parameters as a prior when estimating Naive Bayes parameters using MAP. You may need to implement with log probabilities to avoid underflow.

Task: Train your Naive Bayes classifiers on the training set that is given and report training accuracy as well as testing accuracy. Submit your code.