Bayesian Learning

Debrup Chakraborty

To be covered today

- Estimation of the probabilities
- Naive Bayes Classifier

Materials

- Chapter 3 of Duda
- Chapter 6 of Mitchell

Bayes Desicion Theory

- In the last class we saw that with Bayes Decision theory we can make optimal classifiers
- But Bayes theorem needs the knowledge of the prior probabilities and the class conditional probabilities.
- In practice obtaining those probabilities are quite difficult.
- Today we shall learn about some techniques to estimate the probabilities from data
- Also we shall see how Bayes theory can be used in case of discrete features.

Parameter Estimation

- To apply Bayes theory we need two probabilities. The prior probabilities and the class conditional probabilities.
- Estimating the prior probabilities are quite easy, but obtaining an estimate for the class conditional probabilities from the training data is difficult. This is because in most classification problems the amount of available data is always small and the problems are generally posed in very high dimensions.

Parameter Estimation (Contd.)

- However if by some means we can parametrize the class conditional probabilities, we can estimate the parameters and thus have an estimate of the probabilities.
- Like if we know, that the class conditional probabilities are normals then we can estimate the parameters for the normal density.
- Two schools of thought for parameter estimation
  - Maximum Likelihood Estimation
  - Bayesian Parameter Estimation

Parameter Estimation: MLE

- We assume we have a training set $\mathcal{T}$ which represnts k classes.
- Assume

$$\mathcal{T} = \bigcup_{i=1}^{k} \mathcal{T}_i$$

The data points from the i^{th} class are in $\mathcal{T}_i$.

- Also we assume that the data points in $\mathcal{T}_i$ are independent of that in $\mathcal{T}_j$, where $i \neq j$.
- Assume that points in each $\mathcal{T}_i$ has been independently and identically sampled from some normal distribution with unknown parameters.

Parameter Estimation: MLE

- For simplicity we assume the class conditional probabilities as univariate normals and derive the MLE. But a straight forward gereralization is possible (with some extra computations) to the multivariate case.
- Consider the set $\mathcal{T}_i$. With the points in $\mathcal{T}_i$ we wish to compute the class conditional density $p(x|c_i)$.
- Also it is assumed that the points in $\mathcal{T}_i$ are IID and $p(x|c_i) \sim N(\mu, \sigma)$.
- We compute the following probability:

$$p(\mathcal{T}_i|\boldsymbol{\theta})$$

- This is the likelihood of the data, and we need to maximize it.

MLE (contd.)

- As points in $\mathcal{T}_i$ are IID thus we can say that

$$\begin{aligned} p(\mathcal{T}_i|\boldsymbol{\theta}) &= \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\theta_2} exp\left[-\frac{1}{2}\left(\frac{x-\theta_1}{\theta_2}\right)^2\right] \end{aligned}$$

MLE (contd.)

- As usual we shall maximize the log-likelihood instead of the likelihood. Thus the function to maximize is:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\ln \left(\frac{1}{\sqrt{2\pi\theta_2}} \right) - \frac{1}{2} \left(\frac{x - \theta_1}{\theta_2} \right)^2 \right]$$

MLE (contd.)

- Maximizing $l(\boldsymbol{\theta})$, we get the MLE estimates of the mean and the variance as

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

MLE (contd.)

- In the multivariate case the estimates would be

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^T$$

Estimating the Prior Probabilities

- As mentioned earlier, estimating the prior probabilities are relatively easy.
- An intuitive estimates for the priors can be as

$$P(c_i) = \frac{|\mathcal{T}_i|}{|\mathcal{T}|}$$

Bayesian Estimation

- There is a second school of thought called the Bayesian Paradigm for estimating parameters.
- Generally the estimates obtained by using the Bayesian techniques are not much different from that obtained through MLE, but it is a different way of looking into things.
- In the bayesian paradigm, the parameter values are not assumed to be fixed as in case of MLE.
- Here it is assumed that the parameters also have a certain prior distribution, and by looking at the data their distribution gets modified.

Naive Bayes Classifier

- A classifier for data with binary features
- Uses many unrealistic assumptions
- But still performs very good for certain classification tasks
  - Document Categorization
  - Spam filtering etc.

Naive Bayes Classifier (contd.)

- Let us consider the case of email classification into spam and non spam.
- The features vectors in this case represents whether a word is present in an email or not.
- According to Bayes theory

$$P(c_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|c_i)P(c_i)}{P(\boldsymbol{x})}$$

Naive Bayes Classifier (contd.)

- We make an assumption that the individual words in a document/email are independent (which is ofcourse not true).
- But this simplifies things and thus we can write the class conditional probability as

$$P(\boldsymbol{x}|c_i) = \prod_{j=1}^{p} P(x_j|c_i)$$

- And compute the posterior as

$$P(c_i|\boldsymbol{x}) = P(c_i) \prod_{j=1}^{p} P(x_j|c_i)$$

Computing the probabilities

Let us decide on the following notations

- $n(x_j, c_i)$ = No of e-mails in class c_i containing the word x_j.
- $N(c_i)$ = No of emails in class c_i.
- N_T = Total no of emails.

Thus,

$$P(x_j|c_i) = \frac{n(x_j, c_i)}{N(c_i)}$$

$$P(c_i) = \frac{N(c_i)}{N_T}$$

Computing the probabilities

- This computation of probabilities may sometimes lead to problems.
- A better estimate of the probability:

$$P(x_j|c_i) = \frac{n_{i,j} + 1}{n_i + |\text{Vocabulary}|}$$

$$P(c_i) = \frac{n_i}{N},$$

  - $N =$ Total number of words in all emails
  - $n_i =$ Number of words in emails in class c_i
  - $n_{i,j} =$ Number of times the word x_j occurs in emails with class c_i
  - $|\text{Vocabulary}| =$ Number of words in the vocabulary

MDL Principle

Occam's razor: prefer the shortest hypothesis

MDL: prefer the hypothesis h that minimizes

$$h_{MDL} = \arg\min_{h \in H} L_{C_1}(h) + L_{C_2}(D|h)$$

where $L_C(x)$ is the description length of x under encoding C

MDL Principle

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(D|h)P(h) \\ &= \arg\max_{h \in H} \log_2 P(D|h) + \log_2 P(h) \\ &= \arg\min_{h \in H} - \log_2 P(D|h) - \log_2 P(h) \end{aligned}$$

Interesting fact from information theory:

> The optimal (shortest expected coding length) code for an event with probability p is $-\log_2 p$ bits.

MDL Principle

So interpret :

- $-\log_2 P(h)$ is length of h under optimal code
- $-\log_2 P(D|h)$ is length of D given h under optimal code

→ prefer the hypothesis that minimizes

$$length(h) + length(misclassifications)$$