TECHNO INDIA
SILIGURI INSTITUTE
OF TECHNOLOGY
A Satyam Roychowdhury Initiative

Bayesian Learning

Probability Theory and Bayes Rule, Naïve Bayes Learning Algorithm, Parameter Smoothing, Generative vs Discriminative Training, Bayes Net and Markov Net for representing dependencies

Class Conditional Density

- A general form of classifier for classifying a feature vector x by applying a Bayes rule is

$$P(y = c|x, \theta) \propto P(x|y = c, \theta)P(y = c|\theta)$$

- The key to using such models is specifying a suitable form for the class conditional density $P(x|y = c, \theta)$, which defines what kind of data we expect to see in each class.

Bayesian Concept Learning

- In a concept learning for binary classification problem we define $f(x) = 1$, if x is an example of the concept C and $f(x) = 0$, otherwise. The goal is to learn the indicator function f which just define which elements are in the concept C.
- For example a series randomly chosen positive examples $D = \{x_1, x_2, \ldots, x_n\}$ drawn from C, and ask you whether the new sample $\hat{x}$ belong to C or not i.e. classify $\hat{x}$.
- It can be represented as probability distribution $P(\hat{x}|D)$, probability that $\hat{x} \in C$ given the data D. This is called posterior predictive distribution. We have to guess the concept C of the data D. The classic approach to consider the hypothesis space of concept C.

Likelihood

- Consider the examples are sampled uniformly at random from the extension of concept. Given the assumption the probability of independently sample N items from the hypothesis h is

$$P(D|h) = \left[\frac{1}{|h|}\right]^N$$

Prior

- Some hypothesis are conceptually unnatural. Such intuition can be capture by assigning low probability to the unnatural concepts. This subjective part of Bayesian learning is the source of many controversy.

Posterior

- The posterior is simply the likelihood times of prior.

$$P(h|D) = (P(D|h)P(h))/(\sum_{h'} P(D|h')P(h''))$$

- In case, most of the priors are uniform, so the posterior is proportional to the likelihood.
- In case we have enough data, the posterior becomes peaked for a single concept, namely MAP estimation i.e.

$$P(h|D) \rightarrow \delta_{h'MAP}(h)$$

Where $h'MAP = \arg\max_{\mathrm{h}} P(h|D)$ is the posterior mode and δ is the diarac measure defined by

- $\delta_{x(A)} = 1\ if\ x \in A$, 0 otherwise

Maximum a posteriori estimation

- MAP estimate can be written as

$$h'MAP = \arg\max_{h} P(D|h)P(h) = \arg\max[logP(D|h) + \log P(h)]$$

- Since the likelihood term depends exponentially on N,and the prior stays constants as we get more and more data, the MAP estimate converge towards the maximum likelihood estimate(MLE).

Bayes Theorem

- Bayes theorem: $P(h|D) = \frac{P(D|h)P(h)}{P(D)}$
- If we apply Bayes theorem for classification for classification
- $P(Y|X) \propto P(X|Y)P(Y) \equiv P(x_1, x_2, \dots, x_n|Y)P(Y)$
- Joint probability store probability of different variables for all possible values which is intractable problem.
- For Naïve Bayes, we assume that individual x_i is independent given Y.
- $P(x_1, x_2, \dots, x_n|Y) = P(x_1|Y)P(x_2|x_1, Y) \dots P(x_n|x_1, x_2, \dots, x_{n-1}, Y)$
- $P(x_1, x_2, \dots, x_n|Y) = P(x_1|Y)P(x_2|Y) \dots P(x_n|\, Y)$

Naïve Bayes Classifier

- All the independent variables are conditionally independent to given Y.

$$y^{new} = \arg\max_{k} \prod_{i=1}^{n} P(x_i|Y)P(Y = y_k)$$

Naïve Bayes Classifier-Discrete x_i

- Train Naïve Bayes
  - For each value y_k
  - Estimate $\pi_k = P(Y = y_k)$
  - For each value x_{ij} for each attribute x_i estimate
  - $\theta_{ijk} = P(x_i = x_{ij} | y = y_k)$
- Classify x^{new}
  - $y^{new} = \arg\max_k P(y = y_k) \prod_i P(x_i^{new} | y = y_k)$
  - $y^{new} = \arg\max_k \pi_k \prod_i \theta_{ijk}$