Naïve Bayes

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Problem Setting

- Dataset is a set of possible instances **D** = {(x_1, y_1), (x_2, y_2), …, (x_n, y_n)}
- $\mathbf{x_i} \in \mathbf{X}$**:** Each sample is a vector with $\mathbf{R^d}$ drawn from distribution ***P(x, y)***
- Unknown target function **f : X -> Y** as distribution ***P(y/x)***
- Set of function hypotheses H={ h | h : X->Y }
- **Input:** Training examples {<x_i,y_i>}
- **Output:** Hypothesis h ∈ H that best approximates target function f
- Knowing the P, we can simply bayes theorem to get a perfect hypothesis.
  - $\boldsymbol{P_\theta(x, y) \sim P(x, y)}$
  - We do not have ***P(x, y)*** but we do have **D**

How to get θ?

- **Maximum Likelihood Estimation (MLE)** gives us the solution which maximises the likelihood.
  - Find θ that maximizes the probability of the data D
    - $argmax\ P_\theta(D)$
- **Maximum A Posterior (MAP)** gives us the solution which maximises the posterior probability.
  - Find θ that is most likely given the data D.
    - $P(\theta|D) = P(D|\theta) * P(\theta)/\mathbf{P(D)} \approx P(D|\theta) * P(\theta)$
    - Assumes the availability of the prior $P(\theta) \sim N(0, \sigma_0^2)$
- Both ML and MAP return only single and specific values for the parameter θ!

How to get θ?

- **Bayesian Inference**
  - We are not interested in estimating θ*s*, but in making predictions! $$p_\theta(y|X=x)$$
  - Holy grail of all predictions will average out all possible models one could think of!
    $$p(y|X=x) = \int_\theta p(y|\theta)p(\theta|D)d\theta$$
  - Find θ that is most likely given the data D.
    - $P(\theta|D) = P(D|\theta) * P(\theta)/\mathrm{P(D)}$
  - Very high computational complexity and sample size requirements!!!
    - Hence -> Naive assumption!

Introduction

- Why 'Naïve'?
  - Features are independent of each other
- Rev. Thomas Bayes (1702–61)
  - Existence of God
- Pros:
  - Easy and fast, performs well in multi class prediction
  - Better to other models like logistic regression and need less training data.

Naïve Bayes Model

- Consider each attribute and class label as random variables

- Given a record with attributes $(X_1, X_2, .. X_n)$
  - Goal is to predict class Y
  - Specifically, we want to find the value of Y that maximizes
    - $P(Y = y_1 | X_1, X_2, .. X_n)$
    - $P(Y = y_2 | X_1, X_2, .. X_n)$
    - $P(Y = y_3 | X_1, X_2, .. X_n)$

- Can we estimate $P(Y = y_1 | X_1, X_2, .. X_n)$ directly from data?

Naïve Bayes Model

- Compute the posterior probability $P(Y| X_1, X_2, .. X_n)$ for all values of Y using the Bayes theorem
  - $P(Y| X_1, X_2, .. X_n) =$
    - $P(X_1, X_2, .. X_n |Y) * P(Y)/P(X_1, X_2, .. X_n)$

- Choose value of Y that maximizes $P(Y| X_1, X_2, .. X_n)$
  - Equivalent to choosing value of Y that maximizes $P(X_1, X_2, .. X_n |Y) * P(Y)$

- How to estimate $P(X_1, X_2, .. X_n |Y) * P(Y)$?

Naïve Bayes Model

- Assume independence among attributes X_i when class is given $P(X_1, X_2, .. X_n |Y_j) * P(Y = Y_j) =$
    - $P(Y = Y_j)*P(X_1|Y_j) * P(X_2|Y_j) * .. P(X_n |Y_j)$

- Can estimate $P(X_i| Y_j)$ for all X_i and Y_j.

- New point is classified to Y_j if $P(Y_j) \Pi P(X_i| Y_j)$ is maximum.

Naïve Bayes Classifier

$$Y^{new} = argmax P(Y = Y_j) \prod_i P(X^{new} | Y = Y_j)$$

Parameter Estimation: Discrete (MLE)

- Training in Naïve Bayes is **easy**:
  - Estimate $P(Y=y_j)$ as the fraction of records with $Y=v$

$$P(Y = y_j) = \frac{count(Y = y_j)}{n}$$

  - Estimate $P(X_i=x_{ij}|Y=y_j)$ as the fraction of records with $Y=y_j$ for which $X_i=x_{ij}$

$$P(X = x_{ij}|Y = y_j) = \frac{count(X = x_{ij} \wedge Y = y_j)}{count(X = x_{ij})}$$

Example: P(Red|Yes)

Colour	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

Parameter Estimation: MAP/Smoothing

- If one of the conditional probability is zero, then the entire expression becomes zero
  - *c: number of classes*
  - *p: prior probability*
  - *m: parameter*

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m-estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

Parameter Estimation: Continuous

- Gaussian naive Bayes

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

- Bernoulli naive Bayes

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i} (1 - p_{ki})^{(1-x_i)}$$

- Multinomial naive Bayes

$$p(\mathbf{x} \mid C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i}$$

Parameter Estimation: Numerical

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

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

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

Example: (Red, Domestic, SUV)?

Colour	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

Example: (Red, Domestic, SUV)?

- P(Yes) = 0.5 and P(No) = 0.5
- p = 1 / (number-of-attribute-values) = 0.5 for all of our attributes
- m = 3
- P(Red|Yes), P(SUV|Yes), P(Domestic|Yes) ,
- P(Red|No) , P(SUV|No), and P(Domestic|No)

$$P(Red|Yes) = \frac{3 + 3 * .5}{5 + 3} = .56 \qquad P(Red|No) = \frac{2 + 3 * .5}{5 + 3} = .43$$

$$P(SUV|Yes) = \frac{1 + 3 * .5}{5 + 3} = .31 \qquad P(SUV|No) = \frac{3 + 3 * .5}{5 + 3} = .56$$

$$P(Domestic|Yes) = \frac{2 + 3 * .5}{5 + 3} = .43 \qquad P(Domestic|No) = \frac{3 + 3 * .5}{5 + 3} = .56$$

Summary

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as Bayesian Belief Networks (BBN)