THEORY

OF

NAIVE BAYES

This page intentionally left blank

Learning classifiers by learning $P(Y|X)$

Simple example:

Joint Distributions

X_1	X_2	Y	Prob.
0	0	0	p_1
0	0	1	p_2
0	1	0	p_3
0	1	1	p_4
1	0	0	p_5
1	0	1	p_6
1	1	0	p_7
1	1	1	p_8
			$\sum_i p_i = 1$

X_1	X_2	$P(Y\|X_1,X_2)$
0	0	p_1
0	1	p_2
1	0	p_3
1	1	p_4

Only 4 parameters!

In general, $X = \langle x_1, \dots x_n \rangle \quad x_i \in \{0,1\}$ and $Y \in \{0,1\}$

Q.1. To estimate $P(Y|X_1, X_2 \dots X_n)$, how many parameters? 2^n

How can we design a learning algorithm that is practical?

Let's go back to Bayes Rule and see if it helps

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

which is shorthand for

$$(\forall i,j) \quad P(Y=y_i \,|\, X=x_j) = \frac{P(X=x_j \,|\, Y=y_i)\, P(Y=y_i)}{P(X=x_j)}$$

equivalently,

$$(\forall i,j) \quad P(Y=y_i \,|\, X=x_j) = \frac{P(X=x_j \,|\, Y=y_i)\, P(Y=y_i)}{\sum_K P(X=x_j \,|\, Y=y_K)\, P(Y=y_K)}$$

Q.2. To estimate $P(X_1, X_2, \dots X_n | Y)$, how many parametters?

case1 : $Y = 1$ $\quad P(X_1, X_2 \dots X_n | Y=1) \rightarrow 2^n - 1$

case 2 : $Y = \phi$ $\quad P(X_1, X_2, \dots X_n | Y=0) \rightarrow 2^n - 1$

$$2(2^n - 1)$$

Q. 3. To estimate $P(Y)$, how many parameters? $\quad 1$

So, if we use Bayes Rule, we need $2(2^n - 1) + 1$ parameters (worse!)

Naïve Bayes

Naïve Bayes assumes :

$$P(X_1, \dots X_n | Y) = \prod_i P(X_i | Y)$$

i.e. that X_i and X_j are conditionally independent given Y, $\forall i \neq j$

Conditional Independence

Def. X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

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

shorthand:

$$P(X | Y, Z) = P(X | Z)$$

e.g. $P(\text{Thunder} | \text{Rain}, \text{Lightning}) = P(\text{Thunder} | \text{Lightning})$

NOTE : Thunder and Rain are not independent!

Naïve Bayes uses assumption that X_i are conditionally independent, given Y

e.g. $P(X_1 | X_2, Y) = P(X_1 | Y)$

Given this assumption, then:

$$P(X_1, X_2 | Y) = P(X_1 | X_2 Y)\ P(X_2 | Y) \quad \text{chain rule}$$
$$= P(X_1 | Y)\ P(X_2 | Y) \quad \text{conditional independence}$$

In general, $P(X_1 \dots X_n | Y) = \prod_i P(X_i | Y)$

Q.4 How many parameters to ~~de~~ estimate $P(X_1, \dots X_n | Y)$? $P(Y)$?

- without conditional independence assumption? $2(2^n - 1) + 1$
- with conditional independence assumption? $2n + 1$

<u>Naïve Bayes in a Nutshell</u>

Bayes Rule:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k)\ P(X_1 \dots X_n | Y = y_k)}{\sum_j P(Y = y_j)\ P(X_1 \dots X_n | Y = y_j)}$$

Assuming conditional independence among X_i's

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

So, to pick most probable Y for $X^{new} = \langle X_1, \dots X_n \rangle$

$$Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Naïve Bayes Algorithm for discrete X_i

- Train Naïve Bayes (examples)

  for each* value y_K

  estimate $\pi_K \equiv P(Y = y_K)$

  for each* value x_{ij} of each attribute X_i

  estimate $\theta_{ijK} \equiv P(X_i = x_{ij} \mid Y = y_K)$

* prob. must sum to 1, so we need to estimate only $n-1$ of these ...

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_K} P(Y = y_K) \prod_i P(X_i^{new} \mid Y = y_K)$$

$$Y^{new} \leftarrow \arg\max_{y_K} \pi_K \prod_i \theta_{ijK}$$

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates (MLE's)

$$\hat{\pi}_K = \hat{P}(Y = y_K) = \frac{\#D\{Y = y_K\}}{|D|}$$

$$\hat{\theta}_{ijK} = \hat{P}(X_i = x_{ij} \mid Y = y_K) = \frac{\#D\{X_i = x_{ij} \text{ and } Y = y_K\}}{\#D\{Y = y_K\}}$$

Number of data points for which $Y = y_K$

~~Naive Bayes~~: MAP estimates (Beta, Dirichlet priors):

$$\hat{\pi}_K = \hat{P}(Y = y_K) = \frac{\#D\{Y = y_K\} + (\beta_K - 1)}{|D| + \sum_m (\beta_m - 1)}$$

$$\hat{\theta}_{ijK} = \hat{P}(X_i = x_{ij} \mid Y = y_K) = \frac{\#D\{X_i = x_{ij} \text{ and } Y = y_K\} + (\beta_K - 1)}{\#D\{Y = y_K\} + \sum_m (\beta_m - 1)}$$

Naïve Bayes : ~~solved~~ Issue #1

- Often X_i's are not really conditionally independent.
  - We use Naïve Bayes anyway, and it works "pretty well"
    - resulting in right classification, but not righ prob. [Domingos & Pazzani 1996]

- What is the effect on estimated $P(Y|X)$?
  - Extreme case : what if we have two copies $X_i = X_k$

$$P(Y=1|X) = P(Y=1)\, P(X_1|Y=1) P(X_2|Y=1) \dots$$

Naïve Bayes : Issue #2

If unlucky, the MLE estimate for $P(X_i|Y)$ might be zero.

- Why worry about just one parameter out of many?

$$P(Y|X_1 \dots X_n) = \frac{P(Y=1) \prod_i P(X_i|Y=1)}{P(X_1 \dots X_n)}$$

- What can be done to address it?

  use MAP estimates by using a prior.

Another way to view Naïve Bayes (Bollean Y) : shape of the decision surface

Decision rule: is this quantity greater than or less than 1?

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$1 \gtrless \frac{P(Y=1|X_1 \dots X_n)}{P(Y=0|X_1 \dots X_n)} = \frac{P(Y=1) \prod_i P(X_i|Y=1)}{P(Y=0) \prod_i P(X_i|Y=0)}$$

$$0 \gtrless \ln \frac{P(Y=1|X_1 \dots X_n)}{P(Y=0|X_1 \dots X_n)} = \ln \frac{P(Y=1)}{P(Y=0)} + \sum_i \ln \frac{P(X_i|Y=1)}{P(X_i|Y=0)}$$

- linear sum of a prior term and conditional prob. terms
- if $X_i \in \{0, 1\}$, then decision is a linear function of X_i's

What if we have continious X_i ?

e.g. image classification : X_i is real-valued i^{th} pixel

Naïve Bayes requires $P(X_i \mid Y = y_k)$, but X_i is real-valued (continious)

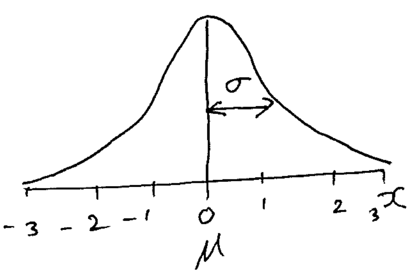
$$P(Y = y_k \mid X_1 \cdots X_n) = \frac{P(Y = y_k) \prod_i P(X_i \mid Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i \mid Y = y_j)}$$

Let's assume $P(X_i \mid Y = y_k)$ follows a Normal (Gaussian) distribution

Gaussian Distribution (a.k.a. Normal distribution)

$p(x)$ is a probability density function (pdf)

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



integral of $p(x)$ is 1.

μ = Mean ; σ^2 = varience

Gaussian Naïve Bayes (GNB)

$$P(X_i = x \mid Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{1}{2}\left(\frac{x-\mu_{ik}}{\sigma_{ik}}\right)^2}$$

Sometimes assume varience

- is independent of Y (i.e. σ_i)
- or independent of X_i (i.e. σ_k)
- or both (i.e. σ)

Gaussian Naïve Bayes Algorithm : continious X_i, but discrete Y

- Train Naïve Bayes (examples)

  for each value y_K

  estimate $\pi_K \equiv P(Y = y_K)$

  for each attribute X_i, estimate $P(X_i | Y = y_K)$

  class conditional mean μ_{iK}, varience σ_{iK}

- Classify (X^{new})

$$Y^{new} \leftarrow \arg\max_{y_K} P(Y = y_K) \prod_i P(X_i^{new} | Y = y_K)$$

$$Y^{new} \leftarrow \arg\max_{y_K} \pi_K \prod_i \mathcal{N}(X_i^{new}; \mu_{iK}, \sigma_{iK})$$

Estimating Parameters : continious X_i, but discrete Y

MLE :

j th training example

$$\hat{\mu}_{iK} = \frac{1}{\sum_j \delta(Y^j = y_K)} \sum_j X_i^j \, \delta(Y^j = y_K)$$

ith feature

k th class

$\delta() = 1$ if $(Y^j = y_K)$
else $\emptyset$

$$\hat{\sigma}^2_{iK} = \frac{1}{\sum_j \delta(Y^j = y_K)} \sum_j (X_i^j - \hat{\mu}_{iK})^2 \, \delta(Y^j = y_K)$$

MAP :