Bayesian Learning

Part II

Naïve Bayes

Some slides were adapted/taken from various sources, including Prof. Andrew Ng's Coursera Lectures, Stanford University, Prof. Kilian Q. Weinberger's lectures on Machine Learning, Cornell University, Prof. Sudeshna Sarkar's Lecture on Machine Learning, IIT Kharagpur, Prof. Bing Liu's lecture, University of Illinois at Chicago (UIC), CS231n: Convolutional Neural Networks for Visual Recognition lectures, Stanford University and many more. We thankfully acknowledge them. Students are requested to use this material for their study only and **NOT** to distribute it.

Naïve Bayes Classifier

- From Bayes theorem: $$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$
- $P(Y|\bar{X}) \propto P(\bar{X}|Y)P(Y)$ where $\bar{X}$ is attributes of input data X

  $= P(X_1X_2X_3 \ldots X_n|Y)\, P(Y)$ where $P(X_1X_2X_3 \ldots X_n|Y)$ joint conditional probability where n is no. of features
- Calculation of joint probability is intractable as for even Boolean valued feature we have to calculate 2^n no. of probability values.
- Naïve Bayes assumption: Attributes that describe instances are conditionally independent given classification

  $= P(X_1|Y)P(X_2|Y)P(X_3|Y)\ldots P(X_n|Y)\, P(Y)$
- So we are assuming the conditional independence among the individual attributes $X_1X_2X_3 \ldots X_n$ and based on this, we can do the classification.
- So all the input features are conditionally independent.

Naïve Bayes Classifier

- Assume target function $f: X \rightarrow V$, where each instance x described by attributed $(a_1, a_2, \ldots, a_n)$ and V is the target class.
- Most probable value of $f(x)$ is:

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2, \ldots, a_n)$$

$$= \arg\max_{v_j \in V} \frac{P(a_1, a_2, \ldots, a_n \mid v_j) P(v_j)}{P(a_1, a_2, \ldots, a_n)}$$

$$= \arg\max_{v_j \in V} P(a_1, a_2, \ldots, a_n \mid v_j) P(v_j)$$

Naïve Bayes assumption: $P(a_1, a_2, \ldots, a_n \mid v_j) = P(a_1|v_j)P(a_2|v_j)\ldots\ P(a_n|v_j)$

$$\prod_i P(a_i \mid v_j)$$

which gives

Naïve Bayes classifier: $$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i \mid v_j)$$

Naïve Bayes Classifier

- Bayes Rule:

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

- Assuming conditional independence among X_i's

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

- So classification rule for $X^{new} = < X_1 \ldots X_n >$ is

$$Y^{new} \leftarrow \underset{y_k}{argmax} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

Naïve Bayes Classifier

- Suppose Y takes 2 values: + and –
- We need to know for all such cases Y=+ and Y=-
- For input features we have to know $P(X_i|Y=+)$, $P(X_i|Y=-)$
- Now if X are three dimensional $\{x_1,x_2,x_3\}$, then we have to calculate

$$P(X_i=x_1|Y=+),\ P(X_i=x_2|Y=+),\ P(X_i=x_3|Y=+)$$

$$P(X_i=x_1|Y=-),\ P(X_i=x_2|Y=-),\ P(X_i=x_3|Y=-)$$

Naïve Bayes Algorithm (Discrete X_i)

- Train Naïve Bayes (Example)
- For each* value y_k

  estimate $\pi_k \equiv P(Y = y_k)$ {prior probability}

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

  estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

- Classify (X^{new})

$$Y^{new} \leftarrow \underset{y_k}{argmax}\, P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \underset{y_k}{argmax}\, \pi_k \prod_i \theta_{ijk}$$

- Probabilities must sum to 1, so need estimate only n-1 parameters

Estimating parameters: Y, X_i discrete valued

- Maximum Likelihood Estimates (MLE's)

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

No. of items in set D for which $Y = y_k$

Estimating parameters: Y, X_i discrete valued

- If unlucky, our MLE estimate for $P(X_i|Y)$ may be zero

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

- MAP estimates

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + l}{|D| + lR}$$

Only difference: "imaginary" examples

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\} + l}{\#D\{Y = y_k\} + lM}$$

Example: Play Tennis

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Example: Naïve Bayes

- Learning Phase:

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

P(Play = Yes) = 9/14 P(Play = No) = 5/14

Example

- Test Phase:
  - Given a new instance, predict its label

    $\mathbf{x}'$=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)
  - Look up tables achieved in the learning phrase

    $P(\text{Outlook}=Sunny \mid \text{Play}=Yes) = 2/9$

    $P(\text{Temperature}=Cool \mid \text{Play}=Yes) = 3/9$

    $P(\text{Huminity}=High \mid \text{Play}=Yes) = 3/9$

    $P(\text{Wind}=Strong \mid \text{Play}=Yes) = 3/9$

    $P(\text{Play}=Yes) = 9/14$

    $P(\text{Outlook}=Sunny \mid \text{Play}=No) = 3/5$

    $P(\text{Temperature}=Cool \mid \text{Play}==No) = 1/5$

    $P(\text{Huminity}=High \mid \text{Play}=No) = 4/5$

    $P(\text{Wind}=Strong \mid \text{Play}=No) = 3/5$

    $P(\text{Play}=No) = 5/14$
  - Decision making with the MAP rule

$$P(Yes \mid \mathbf{x}') \approx [P(Sunny \mid Yes)P(Cool \mid Yes)P(High \mid Yes)P(Strong \mid Yes)]P(\text{Play}=Yes) = 0.0053$$

$$P(No \mid \mathbf{x}') \approx [P(Sunny \mid \text{No})P(Cool \mid \text{No})P(High \mid No)P(Strong \mid No)]P(\text{Play}=No) = 0.0206$$

Given the fact $P(Yes \mid \mathbf{x}') < P(No \mid \mathbf{x}')$, we label $\mathbf{x}'$ to be "*No*".

Naïve Bayes: Assumption of Conditional Independence

- Often X_i are not really conditionally independent

- We can use Naïve Bayes in many cases anyways
  - Surprisingly, often the right classification, even when not the right probability

Gaussian Naïve Bayes (Continuous X)

- Algorithm: Continuous values features
  - Conditional probabilities are often modeled with Normal (Gaussian) distribution

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

  - Sometime assume variance
    - is independent of Y (i.e. σ_i),
    - or independent of (i.e. σ_k),
    - or both (i.e. σ)

Naïve Bayes Algorithm (Continuous X_i)

But still discrete Y

- Train Naïve Bayes (Example)
- For each value y_k

  estimate* $\pi_k \equiv P(Y = y_k)$

  for each attribute X_i, estimate

  class conditional mean μ_{ik} variance σ_{ik}
- Classify (X^{new})

$$Y^{new} <- \underset{y_k}{argmax}\, P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} <- \underset{y_k}{argmax}\, \pi_k \prod_i Normal(X_i^{new}, \mu_{ik}, \sigma_{ik})$$

- Probabilities must sum to 1, so need estimate only n-1 parameters

Estimating Parameters: Y discrete, X_i continuous

Maximum likelihood estimates:

jth training example

$$\widehat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k)$$

ith feature

kth class

δ(z)=1 if z true, else 0

$$\widehat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \widehat{\mu}_{ik})^2 \delta(Y^j = y_k)$$

Example:

Example: Continuous-valued Features

– Temperature is naturally of continuous value.

Yes: 25.2, 19.3, 18.5, 21.7, 20.1, 24.3, 22.8, 23.1, 19.8

No: 27.3, 30.1, 17.4, 29.5, 15.1

– Estimate mean and variance for each class

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n, \quad \sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2$$

$$\mu_{Yes} = 21.64, \ \sigma_{Yes} = 2.35$$
$$\mu_{No} = 23.88, \ \sigma_{No} = 7.09$$

– **Learning Phase**: output two Gaussian models for P(temp|C)

$$\hat{P}(x \mid Yes) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{2 \times 2.35^2}\right) = \frac{1}{2.35\sqrt{2\pi}} \exp\left(-\frac{(x-21.64)^2}{11.09}\right)$$

$$\hat{P}(x \mid No) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{2 \times 7.09^2}\right) = \frac{1}{7.09\sqrt{2\pi}} \exp\left(-\frac{(x-23.88)^2}{50.25}\right)$$

to continue...