Introduction to Machine Learning

Lecture 5 Supervised Learning

- Naïve Bayes Classifier

- Readings: Mitchell Ch 6.9-6.10; Murphy Ch 3.5

- Some materials are courtesy of Vibhave Gogate and Tom Mitchell.
- All pictures belong to their creators.

Supervised Learning

- Given some training examples $< x, f(x) >$ and an unknown function f.
- Find a good approximation of f.

- Step 1. Select the features of x to be used.
- Step 2. Select a hypothesis space H: a set of candidate functions to approximate f.
- Step 3. Select a measure to evaluate the functions in H.
- Step 4. Use a machine learning algorithm to find the best function in H according to your measure.

- **Bayesian Learning: the most probable classification**
- **Naïve Bayes Classifier: additional assumption.**

Optimal Bayes Classifier

- Instance space X: each $x \in$ X is characterized by attributes $(x_1, \dots, x_k)$
- Classifications V.

Sky	Water	Forecast	EnjoySport
Sunny	Warm	Same	Yes

- Underlying truth: a distribution $\Pr[x, v]$ over (x, v).

Most probable classification = $\text{argmax}_{v \in V} \Pr[v|x]$

Optimal Bayes Classifier

- Settings: learn a function $X \to V$
- Instance space X: each $x \in$ X is characterized by attributes $(x_1, \dots, x_k)$
- Classifications V.

Sky	Water	Forecast	EnjoySport
Sunny	Warm	Same	Yes

$$\text{argmax}_{v \in V} \Pr[v|x] = \text{argmax}_{v \in V} \sum_{h \in H \text{ and } h(x)=v} \Pr[h]$$

The probability that x is classified as v.
e.g.
x=(Sunny, Warm, Same)
$\Pr[yes|x]$ is the probability that x is classified as yes.

Optimal Bayes Classifier

- Settings: learn a function $X \to V$
- Instance space X: each $x \in$ X is characterized by attributes $(x_1, \dots, x_k)$
- Classifications V.

Sky	Water	Forecast	EnjoySport
Sunny	Warm	Same	Yes

$$\text{argmax}_{v \in V} \Pr[v|x] = \text{argmax}_{v \in V} \sum_{h \in H \text{ and } h(x)=v} \Pr[h]$$

The distribution over H. This is hard to compute.
We need to estimate $\Pr[h]$ by $\Pr[h|D]$ for each h in H.
The example in Lec 4 is very simple.

Optimal Bayes Classifier

- Settings: learn a function $X \to V$
- Instance space X: each $x \in$ X is characterized by attributes $(x_1, \dots, x_k)$
- Classifications V.

Sky	Water	Forecast	EnjoySport
Sunny	Warm	Same	Yes

$$\text{argmax}_{v \in V} \Pr[v|x] = \text{argmax}_{v \in V} \sum_{h \in H \text{ and } h(x)=v} \Pr[h]$$

Can we compute this directly?
Let us try it. (next page)

Naïve Bayes Classifier

$$\Pr[v|x] = \frac{\Pr[x|v] \Pr[v]}{\Pr[x]}.$$ Bayes theorem.

$$\text{argmax} \Pr[v|x] = \text{argmax} \Pr[x|v]\Pr[v] = \text{argmax} \Pr[x_1, x_2, \dots, x_k|v]\Pr[v]$$

$$x = (x_1, \dots, x_k)$$

Sky	EnjoySport
Sunny	Yes
Sunny	No
Rainy	No
Rain	Yes
Sunny	No

Can you classify a new instance (Sunny)?

$$\Pr[\text{Sunny}|yes, D] \Pr[yes|D] = \frac{1}{2} * \frac{2}{5}$$

$$\Pr[\text{Sunny}|no, D] \Pr[no|D] = \frac{2}{3} * \frac{3}{5}$$

Naïve Bayes Classifier

$$\Pr[v|x] = \frac{\Pr[x|v] \Pr[v]}{\Pr[x]}.$$ Bayes theorem.

$$\text{argmax} \Pr[v|x] = \text{argmax} \Pr[x|v]\Pr[v] = \text{argmax} \Pr[x_1, x_2, \dots, x_k|v]\Pr[v]$$
$$x = (x_1, \dots, x_k)$$

Suppose x_i and y are binary.
There are totally $\Theta(2^k)$ different parameters (instance, classification).

A lot of data needed.

Naïve Bayes Classifier

$\Pr[v|x] = \frac{\Pr[x|v]\Pr[v]}{\Pr[x]}$ Bayes theorem.

argmax $\Pr[v|x]$ = argmax $\Pr[x|v]\Pr[v]$ = argmax $\Pr[x_1, x_2, \ldots, x_k|v]\Pr[v]$
$x = (x_1, \ldots, x_k)$

Suppose x_i and y are binary.
There are totally $\Theta(2^k)$ different parameters (instance, classification).

Sky	Water	Forecast	EnjoySport
Sunny	Warm	Same	Yes
Sunny	Warm	Same	No
Rainy	Warm	Change	No
Sunny	Cool	Change	Yes
Sunny	Cool	Same	no

A lot of data needed.

Can you classify a new instance (Sunny, Warm, Change)?
Pr[Sunny, Warm, Change$|yes, D$]=0
Pr[Sunny, Warm, Change$|no, D$]=0

Naïve Bayes Classifier

argmax $\Pr[v|x]$ = argmax $\Pr[x|v]\Pr[v]$ = argmax $\Pr[x_1, x_2, \ldots, x_k|v]\Pr[v]$
$x = (x_1, \ldots, x_k)$

Suppose x_i and y are binary.
There are totally $\Theta(2^k)$ different parameters (instance, classification).

Naïve Bayes Assumption: **given** the classification v, the attributes are **independent**.

$\Pr[x_1, x_2|v] = \Pr[x_1|v]\Pr[x_2|v]$

Naïve Bayes Classifier

Naïve Bayes Assumption: **given** the classification v, the attributes are **independent**.

$\Pr[x_1, x_2|v] = \Pr[x_1|v]\Pr[x_2|v]$

Sky	Water	Forecast	EnjoySport
Sunny	Warm	Same	Yes

Sky	Water	Forecast	EnjoySport
Sunny	Warm	Same	Yes
Sunny	Warm	Same	No
Rainy	Warm	Change	No
Sunny	Cool	Change	Yes
Sunny	Cool	Same	no

Can you classify a new instance (Sunny, Warm, Change)?

Pr[Sunny, Warm, Change$|yes$]

= Pr[Sunny$|yes$] $*$ Pr[Warm$|yes$]* Pr[Change$|yes$]

Naïve Bayes Classifier

argmax $\Pr[v|x]$ = argmax $\Pr[x|v]\Pr[v]$
= argmax $\Pr[x_1, x_2, \ldots, x_k|v]\Pr[v]$

Naïve Bayes Assumption: **given** the classification v, the attributes are **independent**.

$\Pr[x_1, x_2|v] = \Pr[x_1|v]\Pr[x_2|v]$

- argmax $\Pr[x_1, x_2, \ldots, x_k|v]\Pr[v]$ = argmax $\prod \Pr[x_i|v]\ \Pr[v]$
- Naïve Bayes Classifier: argmax $\prod \Pr[x_i|v]\Pr[v]$

How many parameters now? $O(k)$

Naïve Bayes Classifier

- argmax $\Pr[x_1, x_2, \ldots, x_k|v]\Pr[v]$ = argmax $\prod \Pr[x_i|v]\ \Pr[v]$
- Naïve Bayes Classifier: argmax $\prod \Pr[x_i|v]\Pr[v]$

Can you classify a new instance (Sunny, Warm, Change)?

Pr[Sunny$|yes$] $*$ Pr[Warm$|yes$]* Pr[Change$|yes$]*Pr[yes]

Pr[Sunny$|no$] $*$ Pr[Warm$|no$]* Pr[Change$|no$]*Pr[no]

Naïve Bayes Classifier

- **Naïve Bayes Classifier: argmax** $\prod \Pr[x_i|v]\Pr[v]$
- For a new instance $(x_1, \ldots, x_k)$ and each $v \in V$.
- Estimate $\Pr[x_i|v]$ and $\Pr[v]$ for each x_i and v by data D.

MLE of $\Pr[v]$ and $\Pr[x_i|v]$
Bernoulli Point Estimation.

- $C_{class=v}$: set of instances with classification v in D.
- $C_{X_i=x_i}$: set of instances with x_i as the i-th attribute.

$$\Pr[v]_{MLE} = \frac{|C_{class=v}|}{|D|} \qquad \Pr[x_i|v]_{MLE} = \frac{|C_{class=v} \cap C_{X_i=x_i}|}{|C_{class=v}|}$$

Naïve Bayes Classifier

- **Naïve Bayes Classifier: argmax** $\prod \Pr[x_i|v]\Pr[v]$
- $\Pr[v] = \frac{|C_{class=v}|}{|D|}$, $\Pr[x_i|v] = \frac{|C_{class=v} \cap C_{X_i=x_i}|}{|C_{class=v}|}$
- How to classify (Sunny, Warm, Change)?

Sky	Water	Forecast	EnjoySport
Sunny	Warm	Same	Yes
Sunny	Warm	Same	No
Rainy	Warm	Change	No
Sunny	Cool	Change	Yes
Sunny	Cool	Same	no

$\Pr[yes]$=2/5,
$\Pr[Sunny|yes] = 2/2$,
$\Pr[Warm|yes] = 1/2$
$\Pr[Change|yes] = 1/2$

$\Pr[no]$=3/5,
$\Pr[Sunny|no] = 2/3$,
$\Pr[Warm|no] = 2/3$
$\Pr[Change|no] = 1/3$

Text Classification: Rumor or not Rumor

- Article: a sequence of words.

x= 'I have a red apple and a blue apple'

- $X \to V$
- X: all possible articles
- V={rumor, not rumor}
- D: some labeled articles

v_{opt} =argmax $\Pr[x|v]\Pr[v]$

Text Classification: Rumor or not Rumor

- Text Classification. Rumor or not Rumor??
- Article x: a sequence of words.

x= 'I have a red apple and a blue apple'

- v_{opt} =**argmax** $\Pr[x|v]\Pr[v]$

There are so many sequences of words.

- An article usually has more than 1000 words.
- More than 10,000 common words.
- That is $10{,}000^{1000} = 10^{4000}$
- Atoms in universe: 10^{80}
- Hm..
- We need some assumption/simplification.

Text Classification: Rumor or not Rumor

- Text Classification. Rumor or not Rumor??
- Article x: a sequence of words. (L: length)

$$\text{argmax}_v \Pr[x|v] \Pr[v]$$

x= 'I have a red apple and a blue apple'

- Naïve Bayes Assumption: the positions are independent given the class.

$$\text{argmax}_v \prod_{x=1}^{L} \Pr[X_i = x_i|v] \ \Pr[v]$$

- $\Pr[X_i = x_i|v]$ the probability of observing x_i in the i-th position.

I | have | a | red | apple | and | a | blue | apple

Text Classification: Rumor or not Rumor

- $\Pr[X_i = x_i|v]$ the probability of observing x_i in the i-th position.
- How to compute $\Pr[X_i = x_i|v]$?
- **Positions have the same distribution over the word.**
- $\Pr[\boldsymbol{X_i} = \text{apple}|v] = \Pr[\boldsymbol{X_j} = \text{apple}|v]$
- $\Pr[X_i = \text{apple}|v]$ the probability that the i-th position is 'apple'.
- $\Pr[x_i|v]$ can be computed by $\frac{Count(x_i, v)}{\sum_{x \in X} Count(x,v)}$ where X is the set of all distinct words, and $count(x, v)$ is number of the times word x appears in documents of class v.

Remarks

- **What if part of the new instance is never observed?**
- Suppose you have $x = (x_1, cool\ , x_3)$ but no training instance has Water=Cool.
- We will have $\prod \Pr[x_i|\ yes]\Pr[yes] = \prod \Pr[x_i|\ no]\Pr[no] = 0$ regardless of other attributes of x.

$$\Pr[x_i|v]_{MLE} = \frac{|C_{Y=v} \cap C_{X_i=x_i}|}{|C_{Y=v}|}$$

Sky	Water	Forecast	EnjoySport
Sunny	Warm	Same	Yes
Sunny	Warm	Same	No
Rainy	Warm	Change	No
Sunny	Warm	Change	Yes

- This is not reasonable.

Remarks

- **What if part of the new instance is never observed?**
- Let us use 'smoothing'.

$$\Pr[x_i|v]_{MLE} = \frac{|C_{class=v} \cap C_{X_i=x_i}| + mp}{|C_{class=v}| + m}$$

- m-estimate
- m is a constant and p is the prior estimate of x_i over possible values. If no prior estimate, set $p = 1/k$ where k is the number of possible values of the attribute.
- Pretend you have m extra instances of class v and mp of them have x_i as the attribute value.

Remarks

- **What if part of the new instance is never observed?**
- Suppose you have $x = (x_1, cool\ , x_3)$ but no training instance has Water=Cool.

$$\Pr[x_i|v]_{MLE} = \frac{|C_{class=v} \cap C_{X_i=x_i}| + mp}{|C_{class=v}| + m}$$

{cool, warm}

$$\Pr[cool|yes] = \frac{|C_{Y=v} \cap C_{X_i=x_i}| + mp}{|C_{Y=v}| + m} = \frac{m * 0.5}{2 + m}$$

$$\Pr[cool|no] = \frac{|C_{Y=v} \cap C_{X_i=x_i}| + mp}{|C_{Y=v}| + m} = \frac{m * 0.5}{2 + m}$$

Sky	Water	Forecast	EnjoySport
Sunny	Warm	Same	Yes
Sunny	Warm	Same	No
Rainy	Warm	Change	No
Sunny	Warm	Change	Yes

Example

- A text classification example (from Dan Jurafsky)

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

- We use m-estimate where (m=number of distinct word) and (p=1/m).
- m=6, p=1/6

Pr[j]=1/4
Pr[c]=3/4

Pr[Chinese|j]=(1+1)/(3+6)=2/9
Pr[Tokyo|j]=(1+1)/(3+6)=2/9
Pr[Japan|j]=(1+1)/(3+6)=2/9

Pr[Chinese|c]=(5+1)/(8+6)=3/7
Pr[Tokyo|c]=(0+1)/(8+6)=1/14
Pr[Japan|c]=(0+1)/(8+6)=1/14

- $\Pr[j|x] = \left(\frac{1}{4}\right) * \left(\frac{2}{9}\right)^3 * \left(\frac{2}{9}\right) * \left(\frac{2}{9}\right)$ ≈0.0001
- $\Pr[c|x] = \left(\frac{3}{4}\right) * \left(\frac{3}{7}\right)^3 * \left(\frac{1}{14}\right) * \left(\frac{1}{14}\right)$ ≈0.0003

Remarks

- Naïve Bayes Assumption: the attributes are conditionally independent.
- **What if the assumption is not true?**

$$\Pr[x_1, x_2, \dots, x_k|v] \Pr[v] \neq \prod \Pr[x_i|\ v]$$

- This assumption may not be true, but this approximate performs good.
- A plausible reason: Only need the probability of the correct class to be the largest!
- E.g.
- Truth: $\Pr[f(x) = 1] = 0.99$ and $\Pr[f(x) = 0] = 0.01$. Very likely $f(x) = 1$
- Your estimate: $\Pr[f(x) = 1] = 0.51$ and $\Pr[f(x) = 0] = 0.49$. You say $f(x) = 1$.

Remarks

- Naïve Bayes Assumption: the attributes are conditionally independent.
- **The effect of this assumption:**
- (a) the hypothesis space is restricted to those satisfy this assumption.
- (b) it requires less data (low variance).

$\Pr[x_1, x_2, \dots, x_k|v]$ $\quad$ $\Pr[x_i\ |v]$

(Sky=Sunny, Temperature=Warm, Normal, Strong, Warm, Same)=Yes

- (c) Naïve Bayes classifier finds the most probable classification within this hypothesis.
- If the assumption is true, Naïve Bayes is the optimal classifier.

Dealing with Small Numbers

- A practical issue: $\prod \Pr[x_i|\ v]$
- We are multiplying lots of small numbers: underflow.
- $0.5^{57} = 7\ E\ - 18$
- $\Pr[j|x] = \left(\frac{1}{4}\right) * \left(\frac{2}{9}\right)^3 * \left(\frac{2}{9}\right) * \left(\frac{2}{9}\right)$ ≈0.0001
- $\Pr[c|x] = \left(\frac{3}{4}\right) * \left(\frac{3}{7}\right)^3 * \left(\frac{1}{14}\right) * \left(\frac{1}{14}\right)$ ≈0.0003
- Solution: use log and add.
- $a * b = e^{\ln a + \ln b}$
- Keep the log form.

Gaussian Naïve Bayes

- Continuous features

Sky	Water	Forecast	EnjoySport
Sunny	Warm	Same	Yes

Sky	Water	Forecast	EnjoySport
0.2	0.5	0.7	Yes

$$\text{argmax}_v \prod_{x=1}^{k} \Pr[X_i = x_i|v]\ \Pr[v]$$

- $\Pr[v]_{MLE}$ can be computed in the same way.
- $\Pr[Sky = sunny|yes]_{MLE} = \frac{yes \wedge sunny}{yes}$
- $\Pr[Sky = 0.2|yes]_{MLE} = ??$

Generally, assume a distribution over the domain of Sky.
- Discrete domain {sunny, cloudy}
- Continuous domain R.

- **Assume that** $\Pr[Sky = x|yes]$ **follows a particular distribution.**

Gaussian Naïve Bayes

- Continuous features

Sky	Water	Forecast	EnjoySport
0.2	0.5	0.7	Yes

- Assume $\Pr[Sky = x|yes]$ follows a particular distribution.
- Naïve Bayes + Gaussian
- X_i: the i-th attribute, v_k the k-th classification.
- Assume $\Pr[X_i = x_i|v_k]$ follows $N(\mu_{i,k}, \sigma_{i,k}^2)$

$$\Pr[X_i = x|v_k] = \frac{1}{\sigma_{i,k}\sqrt{2\pi}} e^{\frac{-(x_i-\mu_{i,k})^2}{2\sigma_{i,k}^2}}$$

Point estimation:
Learn $\mu_{i,k}, \sigma_{i,k}^2$ from data

Gaussian Naïve Bayes

- Continuous features
- X_i: the i-th attribute, v_k the k-th class.
- Assume $\Pr[X_i = x_i|v_k]$ follows $N(\mu_{i,k}, \sigma_{i,k}^2)$

Sometimes we assume:
The variance is
Independent of class $\sigma_{i,k}^2 = \sigma_i^2$
Independent of attribute $\sigma_{i,k}^2 = \sigma_k^2$
Independent of both $\sigma_{i,k}^2 = \sigma^2$

$$\Pr[X_i = x|v_k] = \frac{1}{\sigma_{i,k}\sqrt{2\pi}} e^{\frac{-(x_i-\mu_{i,k})^2}{2\sigma_{i,k}^2}}$$

Point estimation:
Learn $\mu_{i,k}, \sigma_{i,k}^2$ from data

Gaussian Naïve Bayes

- Continuous features

$$\Pr[X_i = x_i|v_k] = \frac{1}{\sigma_{i,k}\sqrt{2\pi}} e^{\frac{-(x_i-\mu_{i,k})^2}{2\sigma_{i,k}^2}}$$

$$u_{i,k\,MLE} = \frac{\sum_{x \in class_k} x_i}{|class_k|}$$

$$\sigma_{i,k\,MLE}^2 = \frac{\sum_{x \in class_k}\left(x_i - \mu_{i,k\,MLE}\right)^2}{|class_k|}$$

$class_k$: the set of examples of class k.

Unbiased:
$$\sigma_{i,k\,MLE}^2 = \frac{\sum_{x \in class_k}\left(x_i - \mu_{i,k}\right)^2}{|class_k| - 1}$$

Recap: how to learn a Gaussian.

Summary

- **Naïve Bayes Classifier**

argmax $\Pr[v|x]$

=argmax $\Pr[x|v]\ \Pr[v]$

=argmax $\prod \Pr[x_i\ |v]\ \Pr[v]$

Assume a form for them.
Estimate it by data.

Summary

- **Bayes optimal classifier**
  - Review
- **Naïve Bayes Classifier**
  - What is the assumption? Why we need it?
  - When the attributes are discrete
    - How to estimate the parameters?
    - m-estimate
  - When the attributes are continuous
    - Gaussian naïve Bayes.
    - How to estimate the parameters?

Sky	Water	Forecast	EnjoySport
Sunny	Warm	Same	Yes

- **Text classification**

x= 'I have a red apple and a blue apple'

Group Discussion

- A text classification example (from Dan Jurafsky)

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

- We use m-estimate where (m=number of distinct word) and (p=1/m).
- m=6, p=1/6

Pr[j]=1/4
Pr[c]=3/4

Pr[Chinese|j]=(1+1)/(3+6)=2/9
Pr[Tokyo|j]=(1+1)/(3+6)=2/9
Pr[Japan|j]=(1+1)/(3+6)=2/9

Pr[Chinese|c]=(5+1)/(8+6)=3/7
Pr[Tokyo|c]=(0+1)/(8+6)=1/14
Pr[Japan|c]=(0+1)/(8+6)=1/14

- $\Pr[j|x] = \left(\frac{1}{4}\right) * \left(\frac{2}{9}\right)^3 * \left(\frac{2}{9}\right) * \left(\frac{2}{9}\right)$ ≈0.0001
- $\Pr[c|x] = \left(\frac{3}{4}\right) * \left(\frac{3}{7}\right)^3 * \left(\frac{1}{14}\right) * \left(\frac{1}{14}\right)$ ≈0.0003
- $y = (Japan, Macao, Macao)$