Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

January 26, 2015

Today:
- Bayes Classifiers
- Conditional Independence
- Naïve Bayes

Readings:
Machine Learning (ML),

Ch. 3: Naïve Bayes and Logistic Regression

Two Principles for Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

$$\widehat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

只用估计分子即可

点估计简单，对后验概率估计复杂（无法舍去分母）

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

Maximum Likelihood Estimate

X=1 X=0

P(X=1) = θ

P(X=0) = 1-θ

(Bernoulli)

- Each flip yields boolean value for X

$$X \sim \text{Bernoulli: } P(X) = \theta^X (1-\theta)^{(1-X)}$$

- Data set D of independent, identically distributed (iid) flips produces α_1 ones, α_0 zeros

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Maximum A Posteriori (MAP) Estimate

X=1 X=0

- Data set D of independent, identically distributed (iid) flips produces α_1 ones, α_0 zeros

$$P(D|\theta) = P(\alpha_1, \alpha_0|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

- Assume prior $P(\theta) = Beta(\beta_1, \beta_0) = \frac{1}{B(\beta_1,\beta_0)}\,\theta^{\beta_1-1}(1-\theta)^{\beta_0-1}$
- Then

引入了一个虚拟实验(自己认为的)

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta) = \frac{\alpha_1+\beta_1-1}{(\alpha_1+\beta_1-1)+(\alpha_0+\beta_0-1)}$$

(like MLE, but hallucinating $\beta_1 - 1$ additional heads, $\beta_0 - 1$ additional tails)

Let’s learn classifiers by learning P(Y|X)

Consider Y=Wealth, X=<Gender, HoursWorked>

$$Pr(W=w \mid G=g, H=h) = \frac{Pr(W=w, G=g, H=h)}{\sum_w Pr(G=g, H=h, W=w)}$$

求和求掉了

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

How many parameters must we estimate?

Suppose $X = <X_1, \ldots X_n>$

where X_i and Y are boolean RV's

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

To estimate $P(Y| X_1, X_2, \ldots X_n)$

#parameters: $\frac{2^n+2^n}{2} = 2^n$

共 2^n 种 $X=<X_1,\cdots,X_n>$ Y假定两种 $\begin{cases}0\\1\end{cases}$

If we have 30 boolean X_i's: $P(Y | X_1, X_2, \ldots X_{30})$

$\rightarrow n=30$

$2^{30} \approx 10^9$

Bayes Rule

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

Which is shorthand for:

观测值

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

Equivalently:

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

Can we reduce params using Bayes Rule?

Suppose X =<X_1,… X_n>

where X_i and Y are boolean RV's

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

有一个约束：$\sum_{i=1} x_i = 1$ 故少一个自由度.

How many parameters to define $P(X_1, \ldots X_n \mid Y)$?

分情况讨论：$Y=1 \Rightarrow$ 估计 2^n 个 $(X=\langle X_1, \ldots, X_n\rangle)$

但是 $P(X=x|Y=1) + P(X=x|Y=0)$ 不一定为1 故不可相加凑2,

$\Rightarrow$ 估计 X 需要 $(2^n-1)+(2^n-1) = 2^{n+1}-2$ 个估计.

$\Rightarrow$ 总共要 $(X|Y)$: $(2^{n+1}-2)+1 = 2^{n+1}-1$ 个参数去估计.

How many parameters to define P(Y)?

Naïve Bayes

Naïve Bayes assumes

实际上基本不存在"独立"这种情况

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

估计效果不好但能用

假设：特征相互独立，适用于小数据集.

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

Conditional Independence

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

Definition: 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

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

Which we often write

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

E.g.,

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y. E.g., $P(X_1|X_2,Y) = P(X_1|Y)$

Given this assumption, then:

X_1; X_2 独立

$$P(X_1, X_2|Y) = P(X_1|X_2,Y)\,P(X_2|Y) = P(X_1|Y)\,P(X_2|Y)$$

若诸变量且两两独立； $P(X_1,X_2,X_3|Y) = P(X_1,X_2|\cancel{X_3},Y)\,P(X_3|Y)$

$= P(X_1,X_2|Y)\,P(X_3|Y) = \cdots = P(X_1|Y)\,P(X_2|Y)\cdots P(X_n|Y)$

对于 $\prod_{i=1}^{n} P(X_n|Y)$ ： Y 可取0,1两值.

每个X_n估一次 则共需2n次.

但是！不准确

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y. E.g., $P(X_1|X_2, Y) = P(X_1|Y)$

Given this assumption, then:

$$
\begin{aligned}
P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\
&= P(X_1|Y)P(X_2|Y)
\end{aligned}
$$

in general: $P(X_1 ... X_n|Y) = \prod_i P(X_i|Y)$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y. E.g., $P(X_1|X_2,Y) = P(X_1|Y)$

Given this assumption, then:

$$\begin{aligned} P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

in general: $P(X_1 ... X_n|Y) = \prod_i P(X_i|Y)$

How many parameters to describe $P(X_1 ... X_n|Y)$? $P(Y)$?

- Without conditional indep assumption?
- With conditional indep assumption?

Naïve Bayes in a Nutshell

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

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

Naïve Bayes Algorithm – discrete X_i

- Train Naïve Bayes (examples)

  for each* value y_k 有K个参数(K个类)

  estimate $\pi_k \equiv P(Y = y_k)$ => 估 K个

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

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

  特征 第j个取值 标签

  $i \in 1, n$

  $j \in 1, J$ => nJK 个

  $k \in 1, K$

- Classify (X^{new})

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

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i \theta_{ijk}$$

* probabilities must sum to 1, so need estimate only n-1 of these...

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates (MLE's):

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

统计 y_k 个数

统计全部个数.

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

若样本量很低使某项为0

=> 手动修正

Number of items in dataset D for which $Y=y_k$

目标函数: $l(\theta, \pi) = \ln \Pr(D, \theta, \pi) = \ln C((x_0, y_0) \cdots (x_n, y_n))$

$$= \sum_{i=1}^{n} \ln \Pr(x_i, y_i | \theta, \pi) = \sum_{i=1}^{n} \ln \Pr(x_i | y_i, \theta) P(y_i | \pi)$$

$$= \sum_{i=1}^{n} \ln \Pr(x_i | y_i, \theta) + \sum_{i=1}^{n} \ln \Pr(y_i | \pi)$$

$\frac{\partial l(\pi, \theta)}{} \quad \frac{\partial l_2(\pi)}{}$

$D=\{x_i, y_i\}_{i=1}$ $\frac{\partial f(x,\theta)}{\partial \theta} = \frac{\partial f(\theta)}{\partial \theta} = 0$ $\frac{}{\partial x} = \frac{}{\partial x} = 0$

Naïve Bayes: Subtlety #1

Often the X_i are not really conditionally independent

如果朴素贝叶斯假设不成立，强行使用也可以

- We use Naïve Bayes in many cases anyway, and it often works pretty well
  - often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])

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

⇒ 那么会导致过分关注 X_i

⇒ 学不到

Naïve Bayes: Subtlety #2

If unlucky, our MLE estimate for $P(X_i \mid Y)$ might be zero.
(for example, X_i = *birthdate*. X_i = *Jan_25_1992*)

- Why worry about just one parameter out of many?

  人少的时候会有某些 $p(x_i|y)=0$ 导致最后结果为0.

- What can be done to address this?

  引入先验修正，从 MLE -> MAP.

  特征选择预处理
  或联合概率密度建模

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data $\mathcal{D}$

$$\widehat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given prior probability and the data

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates:

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

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

超参

狄利克雷分布

MAP estimates (Beta, Dirichlet priors):

Only difference:
“imaginary” examples

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_m (\beta_m - 1)}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_j | Y = y_k) = \frac{\#D\{X_i = x_j \wedge Y = y_k\} + (\beta_k - 1)}{\#D\{Y = y_k\} + \sum_m (\beta_m - 1)}$$

What you should know:

- Training and using classifiers based on Bayes rule

- Conditional independence
  - What it is
  - Why it's important

- Naïve Bayes → 由指数集 ⇒ 多项式级(线性集)
  - What it is
  - Why we use it so much
  - Training using MLE, MAP estimates
  - Discrete variables and continuous (Gaussian)

Questions:

- How can we extend Naïve Bayes if just 2 of the X_i's are <u>dependent</u>?

- What does the decision surface of a Naïve Bayes classifier look like?

$$\left\{ x \;\middle|\; \ln \frac{P(Y=y_k \mid X=x)}{P(Y=y_\ell \mid X=x)} = 0 \right\}$$

- What error will the classifier achieve if Naïve Bayes assumption is satisfied and we have infinite training data?

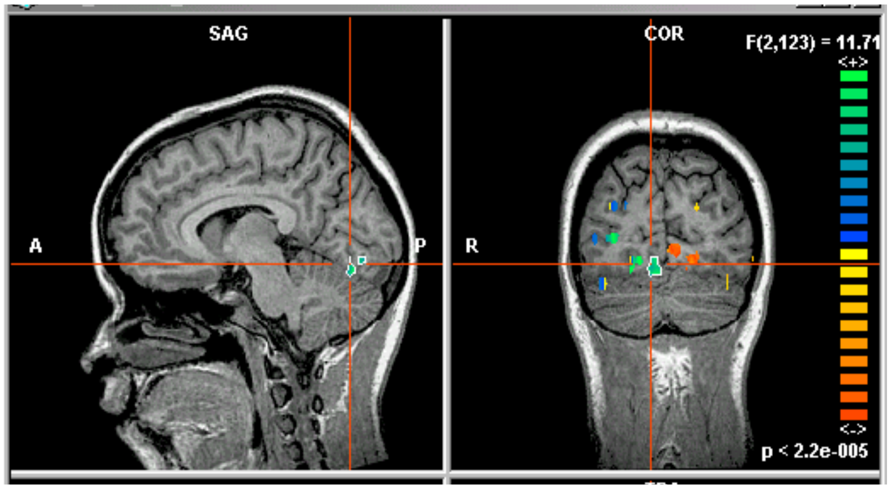
$$\min EPE(f) \Rightarrow \arg\max_{y_k} \Pr(Y=y_k \mid X=x) = \arg\max_{y_k} \frac{\prod_{i=1}^{n} \Pr(X=x_{0,i} \mid Y=y_k) \Pr(Y=y_k)}{\Pr(X=x)} = \pi_k \prod_{ijk} \theta_{ijk}$$

- Can you use Naïve Bayes for a combination of discrete and real-valued X_i?

连续：用 Gaussian distribution 估计 连续 r.v.

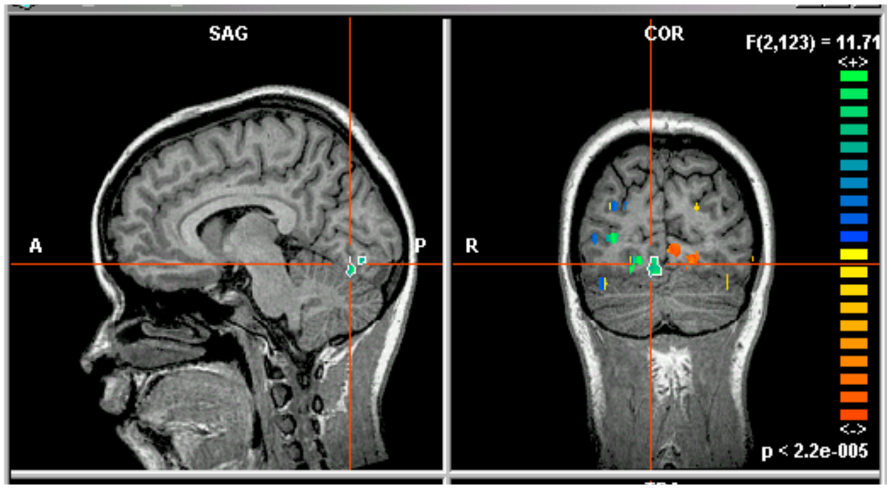
What if we have continuous X_i ?

Eg., image classification: X_i is ith pixel



What if we have continuous X_i ?

image classification: X_i is i^th pixel, Y = mental state



Still have:

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

Just need to decide how to represent $P(X_i \mid Y)$

What if we have continuous X_i ?

Eg., image classification: X_i is ith pixel

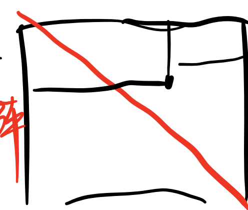
Gaussian Naïve Bayes (GNB): assume

每一个 X_i Y_k 都有自己的一个 σ_{ik}, μ_{ik}

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

一共要估计 2nk 个参数.

GNB: 协方差矩阵 对角矩阵



Sometimes assume σ_{ik}

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

二次绝对边界

线性 (极其简单).

共享一个 σ_k

假设

Diag (Σ_k)

后

是线性边界 ⇐ Diag (Σ) 只剩一个 Σ

Gaussian Naïve Bayes Algorithm – continuous X_i (but still 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

  class conditional mean μ_{ik}, variance σ_{ik}

- Classify (X^{new})

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

$$Y^{new} \leftarrow \arg\max_{y_k} \; \pi_k \prod_i Normal(X_i^{new}, \mu_{ik}, \sigma_{ik})$$

* 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

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

↳ k类别的个数

ith feature

kth class

δ(z)=1 if z true, else 0

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