MLE – LS - MAP

CS385 – Machine Learning - Classification

Review previous

CS385 – Machine Learning

Spam detection – Naïve Bayes

p(Spam)=3/8 Prior

receive a message, not checking it, we believe that with 3/8 probability it is a junk mail

Posterior

After checking, "secret is secret", we believe that with 25/26 probability, by using Naïve Bayes, it is a junk mail

$$P(\text{spam} \mid \text{"secret is secret"})$$

$$= \frac{\text{①}\ \underline{P(\text{"secret"} \mid \text{spam})}\ \underline{P(\text{"is"} \mid \text{spam})} \cdot \underline{P(\text{"secret"} \mid \text{spam})}\ P(\text{spam})}{\text{①} + P(\text{"secret"} \mid \text{Ham}) \cdot P(\text{"is"} \mid \text{Ham}) \cdot P(\text{"secret"} \mid \text{Ham})\ P(\text{Ham})}$$

* Bayes Rules

* Conditional Independent.

X1	⋯	Xn	Y (k classes)

$C_1 \quad C_2 \cdots \quad C_k$

(val1, val2, ⋯ valn)

which class?

$$\max_{C_i} \begin{cases} P(C_1 \mid val_1, val_2, \ldots, val_n) \\ P(C_2 \mid val_1, val_2, \ldots, val_n) \\ \vdots \\ P(C_k \mid val_1, \ldots, val_n) \end{cases}$$

$$P(C_i \mid val_1 \ldots val_n) = \frac{\prod p(val_j \mid C_i)}{P(val_1 \ldots val_n)}$$

Maximum Likelihood Estimate

$\pi =$ **p(Spam)= 3/8** **how do we learn it?**

p('is'|Spam)= 1/15 **how do we learn it?**

$p(\text{data set} \mid \bar{\pi})$

$= p(m_1, m_2, \ldots, m_n \mid \pi)$ $\quad$ $\Sigma \cdot \Sigma 1)$

$= p(m_1 \mid \bar{\pi})\, p(m_2 \mid \bar{\pi}) \ldots p(m_n \mid \bar{\pi})$

$\operatorname{argmax}_{\pi}$

Laplace Smoothing

CS385 – Machine Learning

Overfitting – Laplace Smoothing

P(Spam| "Today is Secret") =0

MLE P(x)= count(x)/N

LS P(x)= (count(x) + k) / (N+k|x|)

K=1 p(spam)= ?

p(x = Spam)

p(x = Ham)

1 message	1 spam	2/3	$= \frac{1+1}{1+1\times2}$
10 messages	6 spam	7/12	$\frac{6+1}{10+1\times2} = \frac{7}{12}$
100 messages	60 spam	61/102	

0

p("Today"|spam) ×

p("is"|spam) ×

p("Secret"|spam) ×

p(spam)

LS
k=1

Spam Detection – Training Data

Spam

- Offer is secret
- Click secret link
- Secret sports link

Ham

- Play sports today
- Went play sports
- Secret sports event
- Sport is today
- Sport costs money

1 Size of vocabulary? 12

2 P(Spam) = ? $\frac{3+1}{8+2} = \frac{4}{10} = \frac{2}{5}$

LS Solutions for conditional probability

LS
k=1

Spam

- Offer is secret
- Click secret link
- Secret sports link

Ham

- Play sports today
- Went play sports
- Secret sports event
- Sport is today
- Sport costs money

p (w = 'secret' | spam)

1 P("Secret" |Spam) = $\frac{3+1}{9+1\times 12} = \frac{4}{21}$

2 P("Secret" |Ham) = $\frac{1+1}{15+1\times 12} = \frac{2}{27}$

LS
k=1

Spam

- Offer is secret
- Click secret link
- Secret sports link

Ham

- Play sports today
- Went play sports
- Secret sports event
- Sport is today
- Sport costs money

P(Spam| "Today is secret")

P("is" |Spam) = $\frac{1+1}{9+12} = \frac{2}{21}$

P("is" |Ham) = $\frac{1+1}{15+12} = \frac{2}{27}$

P("today" |Spam) = $\frac{0+1}{9+12} = \frac{1}{21}$

P("today" |Ham) = $\frac{2+1}{15+12} = \frac{3}{27}$

LS(k=1) for detection

Spam

- Offer is secret
- Click secret link
- Secret sports link

Ham

- Play sports today
- Went play sports
- Secret sports event
- Sport is today
- Sport costs money

P(Spam| “Today is Secret”) =0.4858

LS(k=1) for Conditional Probability

8

3 Ham

5 spam

	Spam	Spam(LS)	Ham	Ham(LS)
offer	1	1	0	1
is	1	1	1	1
Secret	3 +	1	1 +	1
click	1	1	0	1
link	2	1	0	1
Sport	1	1	5	1
Play	0	1	2	1
Today	0	1	2	1
Go	0	1	1	1
Event	0	1	1	1
Cost	0	1	1	1
Money	0	1	1	1
	9 +	12	15 +	12

MAP vs. MLE

CS385 – Machine Learning

Finding **π**

MLE: choose π that maximizes probability of observed data

$$\hat{\pi} = \arg\max_{\pi} P(data \mid \pi)$$

MAP (Maximum A Posterior): choose π that is most probable given prior probability and the data

$$\hat{\pi} = \arg\max_{\pi} P(\pi \mid data)$$

$$= \arg\max_{\pi} \frac{P(data \mid \pi)P(\pi)}{P(data)}$$

MLE – **π**, conditional probability **θ**

$$\hat{\pi} = \arg\max_{\pi} P(data \mid \pi)$$

$\pi = P(spam)$

$$\hat{\pi} = \frac{\# data\{Y = 1\}}{|data|} \quad \frac{3}{8}$$

$$\hat{\theta}_{y=1} = \frac{\# data\{X_i = x_{ij} \wedge Y = 1\}}{\# data\{Y = 1\}} = \frac{3}{9}$$

$\theta = P(word \mid spam)$

$\downarrow$

secret.

Number of items in dataset for which Y=1

MAP for Binomial (1)

$p(\pi/data)$

$$\hat{\pi} = \arg\max_{\pi} \frac{P(data \mid \pi)P(\pi)}{P(data)}$$

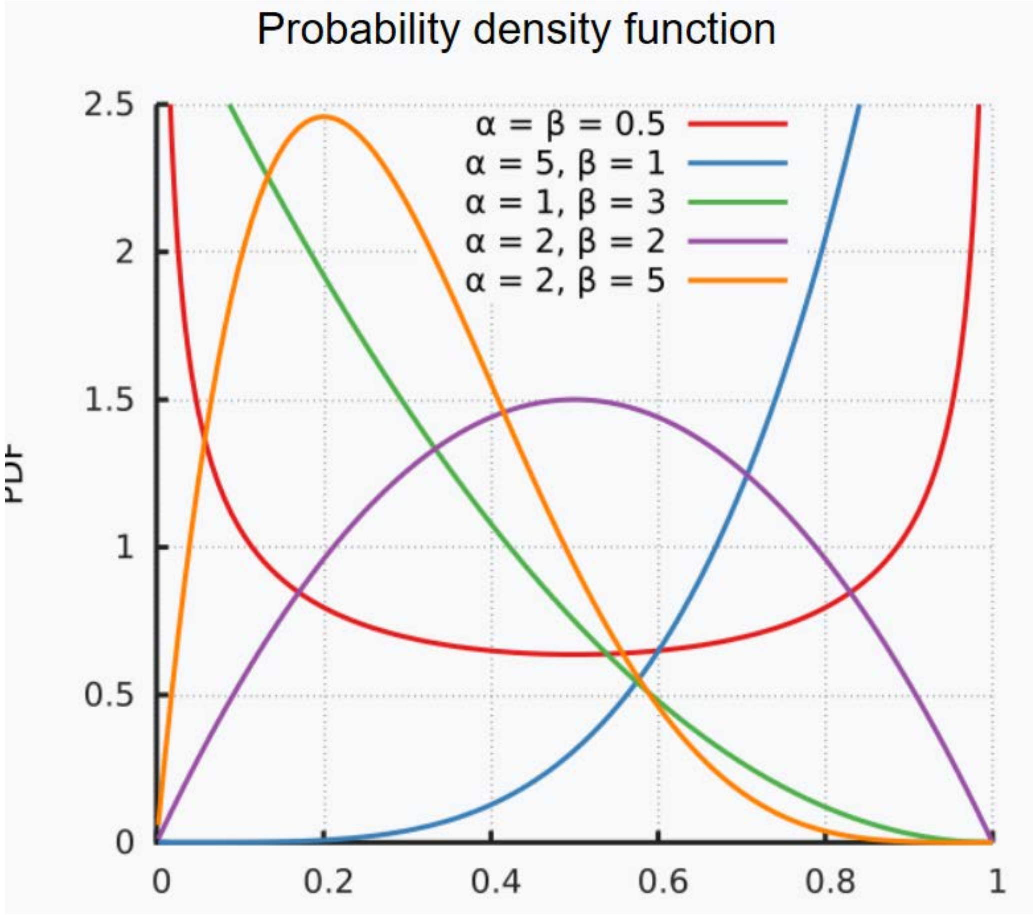
Conjugate priors: P(π) and P(π| data) have the same form, say, Beta distribution for Binomial

Likelihood P(data|π) is Binomial

$$P(data \mid \pi) = \pi^{count(y_i=1)} \cdot (1-\pi)^{count(y_i=0)}$$

$$= \pi^3 \cdot (1-\pi)^5$$

Beta Distribution



Notation	Beta(*α*, *β*)
Parameters	*α* > 0 shape (real) *β* > 0 shape (real)
Support	$x \in [0,1]$ or $x \in (0,1)$
PDF	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$ where $\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

$$
\begin{aligned}
f(x;\alpha,\beta) &= \mathrm{constant}\cdot x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,du} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,x^{\alpha-1}(1-x)^{\beta-1} \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}
\end{aligned}
$$

MAP for Binomial (2)

Prior is Beta distribution

$$P(\pi) = \frac{\pi^{\alpha_1 - 1}(1-\pi)^{\alpha_0 - 1}}{B(\alpha_1, \alpha_0)} \sim Beta(\alpha_1, \alpha_0)$$

Posterior is Beta distribution

$$P(\pi \mid data) = \frac{\pi^{count(y_i=1)} \cdot (1-\pi)^{count(y_i=0)} \cdot \frac{\pi^{\alpha_1 - 1}(1-\pi)^{\alpha_0 - 1}}{B(\alpha_1, \alpha_0)}}{\int_0^1 \pi^{count(y_i=1)} \cdot (1-\pi)^{count(y_i=0)} \cdot \frac{\pi^{\alpha_1 - 1}(1-\pi)^{\alpha_0 - 1}}{B(\alpha_1, \alpha_0)} \, d\pi}$$

$$= \frac{\pi^{count(y_i=1)+\alpha_1-1} \cdot (1-\pi)^{count(y_i=0)+\alpha_0-1}}{\int_0^1 \pi^{count(y_i=1)+\alpha_1-1} \cdot (1-\pi)^{count(y_i=0)+\alpha_0-1}}$$

$$= \frac{\pi^{count(y_i=1)+\alpha_1-1} \cdot (1-\pi)^{count(y_i=0)+\alpha_0-1}}{B(count(y_i = 1) + \alpha_1, count(y_i = 0) + \alpha_0)}$$

$$\sim Beta(count(y_i = 1) + \alpha_1, count(y_i = 0) + \alpha_0)$$

MAP for Binomial (3)

MAP

$3 + \alpha_1$

$$\hat{\pi} = \frac{\# \, data\{Y = 1\} + \alpha_1}{|\, data \,| + \alpha_1 + \alpha_0}$$

0

α_0 α_1 : “Imaginary” Examples

MLE

$$\hat{\pi} = \frac{\# \, data\{Y = 1\}}{|\, data \,|}$$

LS

$$\hat{\pi} = \frac{\# \, data\{Y = 1\} + \mathrm{k}}{|\, data \,| + 2\mathrm{k}}$$

P(word|Spam) Binomial Multinomial ?

- P(spam) P(ham)
- An email is either a spam or a ham

- P(word|spam)
- A word can be “secret”, “today”, “sport”, ….
- A word can be one of the words in the dictionary

- So it follows a multinomial distribution

MAP for multinomial (1)

$\eta \rightarrow P(spam)$ θ_1

$1-\eta \rightarrow P(Ham)$ θ_2

$P(cold)$ ... ...

$\{\theta_1, \theta_2, \theta_3\}$

Dirichlet distribution

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$\theta_1^{\beta_1 - 1} \cdot \theta_2^{\beta_2 - 1} \cdot \theta_3^{\beta_3 - 1}$

$$P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

MAP for multinomial (2)

second spam

$3 + \beta_i$

MAP $$\hat{\theta}_{y=1} = \frac{\# data\{X_i = x_{ij} \wedge Y = 1\} + \beta_{(y=1, x_i = x_{ij})}}{\# data\{Y = 1\} + \sum_m \beta_{(y=1, x_i = x_{im})}}$$

$9 + \beta_1 + \beta_2 + \cdots \beta_{12}.$

MLE

$$\hat{\theta}_{y=1} = \frac{\# data\{X_i = x_{ij} \wedge Y = 1\}}{\# data\{Y = 1\}}$$

LS

$$\hat{\theta}_{y=1} = \frac{\# data\{X_i = x_{ij} \wedge Y = 1\} + k}{\# data\{Y = 1\} + |vocabulary| \cdot k}$$

$\beta_1 = \beta_2 = \cdots \beta_{12} = 1$

Advanced Spam Filters

- Known spamming IP?
- Have you emailed person before?
- Have 1000 other people received same message?
- Email header consistent
- All caps?
- Are you addressed by name?

Gaussian Naïve Bayes

CS385 – Machine Learning

Continuous value

Own Property	Marital Status	Annual Revenue(k)	Delinquency
Y	Single	125	N
N	Married	100	N
N	Single	70	N
Y	Married	120	N
N	Divorced	95	Y
N	Married	60	N
Y	Divorced	220	N
N	Single	85	Y
N	Married	75	N
N	Single	90	Y

Continuous X_i

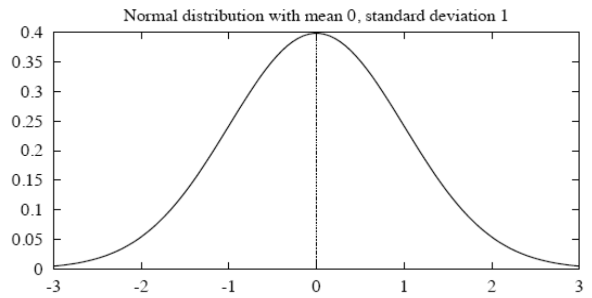
- Y_i is discrete, but X_i is real (continuous)

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

- Common approach
  - Assume $P(X_i|Y=y_k)$ follows a Normal (Gaussian) distribution

Normal (Gaussian) distribution

- p(x) is a probability density function whose integral is 1



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

The probability that X will fall into the interval (a, b) is given by

$$\int_a^b p(x)dx$$

- Expected, or mean value of X, $E[X]$, is

$$E[X] = \mu$$

- Variance of X is

$$Var(X) = \sigma^2$$

- Standard deviation of X, σ_X, is

$$\sigma_X = \sigma$$

GNB - Gaussian Naïve Bayes

- GNB assume $$p(X_i = x|Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{1}{2}(\frac{x-\mu_{ik}}{\sigma_{ik}})^2}$$
- Train Naïve Bayes
  - For each y_k, estimate $\pi_k = p(Y= y_k)$ [how many parameters? n-1, n: number of classes]
  - For each attribute x_i, estimate $p(x_i|Y=y_k)$, class conditional mean μ_{ik} , variance σ_{ik} [how many parameters? 2dn, d: number of attributes]
- Classify(X^{new}) $$Y^{new} \leftarrow \arg\max_{y_k} P(Y = y_k) \prod_i P(X_i^{new}|Y = y_k)$$ $$Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i \mathcal{N}(X_i^{new}; \mu_{ik}, \sigma_{ik})$$

MLE: Y discrete, X_i continuous

- $\pi_k = p(Y = y_k)$
  - Same as Naïve Bayes Discrete
- $p(x_i|Y=y_k)$

jth training example

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

ith feature

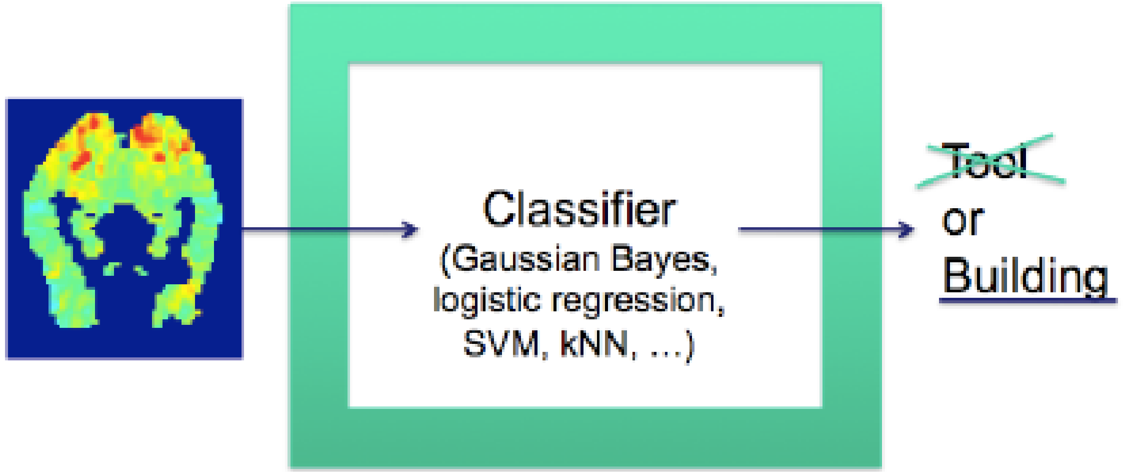
kth class

δ()=1 if ($Y^j=y_k$) else 0

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

GNB Example

- Classify a person's cognitive state based on brain image
- Reading a word describing a "tool" or "building"?



Naïve Bayes - Conclusion

- Designing classifiers based on Bayes rule
- Conditional independence
- How to train Naïve Bayes classifiers
  - MLE and MAP estimates
  - with discrete and/or continuous inputs X_i