Machine Learning
5006.001/2, spring 2016
Session 4: Naive Bayes classifier. Semi-supervised and unsupervised cases through Expectation Maximization algorithm

Instructors: Prof. Stanislav Sobolevsky, Dr. Martin Jankowiak, Dr. Ravi Schroff
Teaching Assistants: Lingjing Wang and Yash Chhajed

Classification

Input/Features Discrete labels Dependence

x_1 y_1

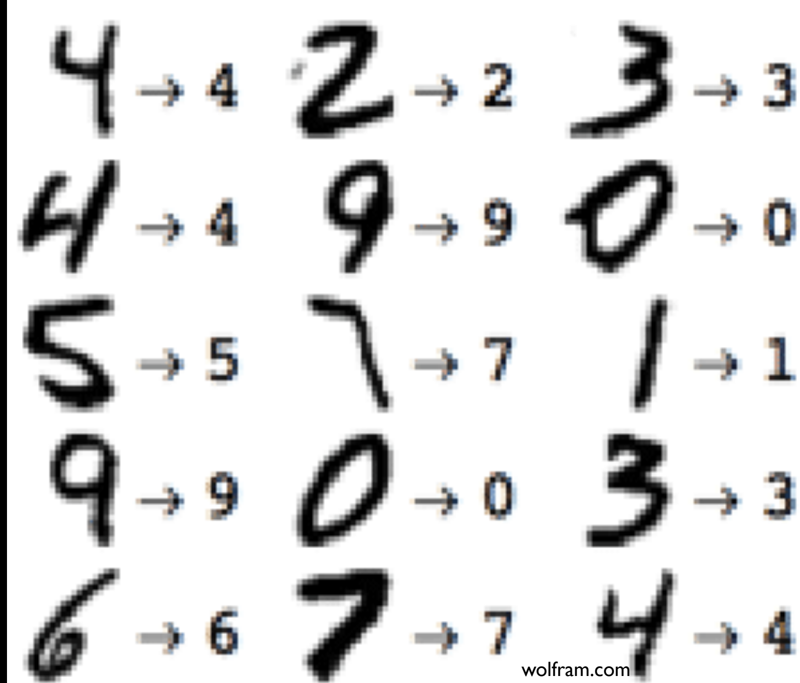
x_2 y_2

$\dots$ $\dots$

x_N y_N

$$y = f(x)$$

$$x^* \quad y^*$$



wolfram.com

Spam classification “toy” example

	Cruise	Lottery	Win	spam
0	1	1	1	1
1	1	0	1	1
2	0	1	0	1
3	0	0	1	0
4	1	0	0	0
5	0	1	0	0

	Cruise	Lottery	Win
0	1	1	0
1	0	0	0

Spam or ham?

Naive Bayes Classifier (discrete case)

$$x = (x^i, i = 1, \ldots, n) \Rightarrow y$$

$$\{(y_j, x_j), j = 1..N\} \qquad x_j = (x_j^i, i = 1, \ldots, n)$$

$$x = x^* \qquad P(y = 0|x = x^*) \qquad P(y = 1|x = x^*)$$

$$P(y = b|x) = \frac{P(x|y = b)P(y = b)}{P(x)}$$

Naive Bayes Classifier (discrete case)

	Cruise	Lottery	Win	spam
0	1	1	1	1
1	1	0	1	1
2	0	1	0	1
3	0	0	1	0
4	1	0	0	0
5	0	1	0	0

$$P(y=b|x)=\frac{P(x|y=b)P(y=b)}{P(x)}$$

$$P(y=b) \qquad P(y=b)=\frac{|\{j: y_j=b\}|}{N}$$

$$P(x=x^*|y=b)=\prod_{i=1}^{n}P(x^i=x^{*i}|y=b)$$ -naive independence assumption

$$P(x^i=x^{*i}|y=b)=\frac{|\{j: x_j^i=x^{*i}, y_j=b\}|}{|\{j: y_j=b\}|}$$

$$P(y=b|x=x^*)\sim P(y=b)\prod_{i=1}^{n}P(x^i=x^{*i}|y=b)$$

$$x=x^* \Rightarrow y^*=\mathrm{argmax}_b P(y=b)\prod_{i=1}^{n}P(x^i=x^{*i}|y=b)$$

Spam classification "toy" example

	Cruise	Lottery	Win	spam
0	1	1	1	1
1	1	0	1	1
2	0	1	0	1
3	0	0	1	0
4	1	0	0	0
5	0	1	0	0

$P(spam) = 1/2$

$P(ham) = 1/2$

$P(Cruise|spam) = 2/3$

$P(Lottery|spam) = 2/3$

$P(Win|spam) = 2/3$

$P(Cruise|ham) = 1/3$

$P(Lottery|ham) = 1/3$

$P(Win|ham) = 1/3$

$$P(spam|\{Cruise, Lottery\}) \sim P(Cruise|spam)P(Lottery|spam)P(!Win|spam)P(spam) =$$
$$=2/3*2/3*(1-2/3)*1/2=2/27$$

$$P(ham|\{Cruise, Lottery\}) \sim P(Cruise|ham)P(Lottery|ham)P(!Win|ham)P(ham) =$$
$$=1/3*1/3*(1-1/3)*1/3=1/27$$

$$P(spam|\{Cruise, Lottery\}) > P(ham|\{Cruise, Lottery\}) \Rightarrow \{Cruise, Lottery\} \rightarrow spam$$

$$P(spam|none) \sim P(!Cruise|spam)P(!Lottery|spam)P(!Win|spam)P(spam) =$$
$$=1/3*1/3*1/3*1/2=1/54$$

$$P(ham|none) \sim P(!Cruise|ham)P(!Lottery|ham)P(!Win|ham)P(ham) =$$
$$=2/3*2/3*2/3*1/2=4/27$$

$$P(spam|\{Win\}) < P(ham|\{Win\}) \Rightarrow \{Win\} \rightarrow ham$$

Example 1

Ipython notebook NYU classes - resources - session4
download the NBsession4.ipynb,
download and unzip data.zip in the same folder

Naive Bayes Classifier (Gaussian continuous case)

$$x = (x^i, i = 1, \ldots, n) \Rightarrow y$$

$$\{(y_j, x_j), j = 1..N\} \qquad x_j = (x_j^i, i = 1, \ldots, n)$$

$$x = x^* \qquad P(y = 0|x = x^*) \qquad P(y = 1|x = x^*)$$

$$P(y = b|x = x^*) \sim p(x = x^*|y = b)P(y = b) = P(y = b)\prod_{i=1}^{n} p(x^i = {x^*}^i|y = b)$$

$$P(y = b) = \frac{|\{j : y_j = b\}|}{N} \qquad \mu_{i,b} = \frac{\sum\limits_{j:x_j^i = {x^*}^i, y_j = b} x_j^i}{|\{j : x_j^i = {x^*}^i, y_j = b\}|}$$

$$p(x^i = {x^*}^i|y = b) = \frac{1}{\sqrt{2\pi}\sigma_{i,b}} e^{-\frac{({x^*}^i - \mu_{i,b})^2}{2\sigma_{i,b}^2}} \qquad \sigma_{i,b} = \sqrt{\frac{\sum\limits_{j:x_j^i = {x^*}^i, y_j = b} (x_j^i - \mu_{i,b})^2}{|\{j : x_j^i = {x^*}^i, y_j = b\}|}}$$

Naive Bayes Classifier (continuous case)

$$P(y=b|x=x^*) \sim p(x=x^*|y=b)P(y=b) = P(y=b)\prod_{i=1}^{n} p(x^i = {x^*}^i|y=b)$$

$$\sim P(y=b)e^{-\sum_{i=1}^{n}\frac{({x^*}^i-\mu_{i,b})^2}{2\sigma_{i,b}^2}}$$

$$y^* = \mathrm{argmax}_b\left[ln(P(y=b)) - \sum_{i=1}^{n}\frac{({x^*}^i-\mu_{i,b})^2}{2\sigma_{i,b}^2}\right]$$

Example 2,3

Ipython notebook NYU classes - resources - session4
download the NBsession4.ipynb,
download and unzip data.zip in the same folder

Bayesian classifier with missing labels: Semi-supervised case

$x = (x^i, i = 1, \ldots, n) \Rightarrow y$

$\{(y_j, x_j), j = 1..N\}$ $\quad x_j = (x_j^i, i = 1, \ldots, n)$

some of the y_j are not known $\quad y_j = nan$

Step 1. Based on the labeled subset of the training data (i.e. having $y_j \neq nan$) estimate sample conditional probabilities $\theta^0_{b,i,a} = P(x^i = a|y = b)$ as well as sample prior probabilities $\theta^0_b = p(y = b)$ for each of the possible values b of y and a of x^i. Set $t = 0$.

Step 2. Given the current estimate $\theta = \theta^t$ for $P(x^i|y = b)$ and $P(y = b)$ define the unobserved labels y_j as the discrete random variables $\hat{y}_j$ with the probability distribution:

$$P(\hat{y}_j = b|x_j, \theta^t) = \frac{P(y = b)\prod_{i=1}^{n} P(x_j^i|y = b)}{\sum_c P(y = c)\prod_{i=1}^{n} P(x_j^i|y = c)} = \frac{\theta_b^t \prod_{i=1}^{n} \theta^t_{b,i,x_j^i}}{\sum_c \theta_c^t \prod_{i=1}^{n} \theta^t_{c,i,x_j^i}}.$$

$$P(\hat{y}_j = b|x_j, \theta^t) = \begin{cases} 1, b = y_j \\ 0, b \neq y_j \end{cases}$$

Bayesian classifier with missing labels: Semi-supervised case

Step 3. Re-estimate the parameters $\theta = \theta^{t+1}$ (maximizing the likelihood of the actual observations, see below) for the distributions of $P(x^i|y = b)$ as well as $P(y = b)$ given the label probabilistic estimates with respect to the probabilities $P(\hat{y}_j = b|x_j, \theta^t)$ defined in step 2.

Step 4. If θ does not change much, i.e. if the termination condition

$$||\theta^{t+1} - \theta^t||_2 = \sum_b (\theta_b^{t+1} - \theta_b^t)^2 + \sum_{k,i,a} (\theta_{k,i,a}^{t+1} - \theta_{k,i,a}^t)^2 < \varepsilon$$

holds - stop with a final estimate $\theta = \theta^{t+1}$, otherwise set $t := t + 1$ and repeat from step 2.

$$\theta_b^{t+1} = P(y = b) = \frac{\sum_j P(\hat{y}_j = b|x^j, \theta^t)}{|\{j\}|}$$

$$\theta_{b,i,a}^{t+1} = P(x_i = a|y = b) = \frac{\sum_{j, x_i^j = a} P(\hat{y}_j = b|x^j, \theta^t)}{\sum_j P(\hat{y}_j = b|x^j, \theta^t)}$$

Bayesian classifier with missing labels: Semi-supervised case

$$\theta^0 \to \theta^1 \to \theta^2 \to \ldots \to \theta^*$$

log-likelihood

$$L(\theta) = \sum_j log\left(\sum_b P(y=b)\prod_i P(x_i = x_i^j | y=b)\right) = \sum_j log\left(\sum_b \theta_b \prod_i \theta_{b,i,x_j^i}\right)$$

$$y^* = \text{argmax}_b P(y=b)\prod_{i=1}^n P(x^i = {x^*}^i | y=b) = \text{argmax}_b \theta_b \prod_{i=1}^n \theta_{b,i,x^{i*}}$$

Example 4

Ipython notebook NYU classes - resources - session4
download the NBsession4.ipynb,
download and unzip data.zip in the same folder

Bayesian unsupervised clustering

$$x = (x^i, i = 1, \ldots, n) \Rightarrow y$$

$$\{(y_j, x_j), j = 1..N\} \qquad x_j = (x_j^i, i = 1, \ldots, n)$$

So what if none of the labels y_j are observed? $y_j = nan$

$$\theta^0 \qquad \sum_k \theta_k^0 = 1 \qquad \sum_a \theta_{k,i,a}^0 = 1$$

$$\theta^0 \to \theta^1 \to \theta^2 \to \ldots \to \theta^*$$

$$y^* = \mathrm{argmax}_b P(y = b) \prod_{i=1}^{n} P(x^i = x^{*i} | y = b) = \mathrm{argmax}_b \theta_b \prod_{i=1}^{n} \theta_{b,i,x^{i*}}$$

Example 5

Ipython notebook NYU classes - resources - session4
download the NBsession4.ipynb,
download and unzip data.zip in the same folder