Machine Learning

4771

Instructor: Tony Jebara

Topic 8

- Discrete Probability Models
- Independence
- Bernoulli Distribution
- Text: Naïve Bayes
- Categorical / Multinomial Distribution
- Text: Bag of Words

Bernoulli Probability Models

- Bernoulli: recall binary (coin flip) probability, just 1x2 table

$$p(x) = \alpha^{x}(1-\alpha)^{1-x} \qquad \alpha \in [0,1] \;\; x \in \{0,1\}$$

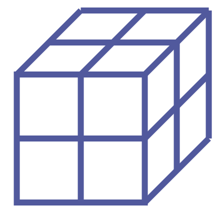
x=0	x=1
0.73	0.27

- Multidimensional Bernoulli: multiple binary events

$$p(x_1, x_2)$$

	x_2=0	x_2=1
x_1=0	0.4	0.1
x_1=1	0.3	0.2

$$p(x_1, x_2, x_3)$$



- Why do we write these as an equations instead of tables?

Bernoulli Probability Models

- Bernoulli: recall binary (coin flip) probability, just 1x2 table

$$p(x) = \alpha^{x}(1-\alpha)^{1-x} \qquad \alpha \in [0,1] \quad x \in \{0,1\}$$

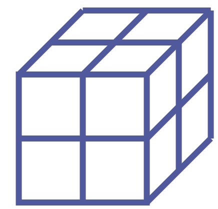
x=0	x=1
0.73	0.27

- Multidimensional Bernoulli: multiple binary events

$$p(x_1, x_2)$$

	x_2=0	x_2=1
x_1=0	0.4	0.1
x_1=1	0.3	0.2

$$p(x_1, x_2, x_3)$$



- Why do we write these as an equations instead of tables?

- To do things like… maximum likelihood…
- Fill in the table so that it matches real data…
- Example: coin flips H,H,T,T,T,H,T,H,H,H ???

x=T	x=H

Bernoulli Probability Models

- Bernoulli: recall binary (coin flip) probability, just 1x2 table

$$p(x) = \alpha^x (1-\alpha)^{1-x} \qquad \alpha \in [0,1] \quad x \in \{0,1\}$$

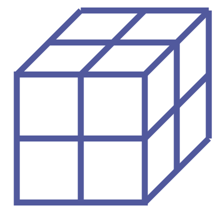
x=0	x=1
0.73	0.27

- Multidimensional Probability Table: multiple binary events

$$p(x_1, x_2)$$

	x_2=0	x_2=1
x_1=0	0.4	0.1
x_1=1	0.3	0.2

$$p(x_1, x_2, x_3)$$



- Why do we write these as an equations instead of tables?

- To do things like… maximum likelihood…
- Fill in the table so that it matches real data…
- Example: coin flips H,H,T,T,T,H,T,H,H,H
- Why is this correct?

x=T	x=H
0.4	0.6

Bernoulli Maximum Likelihood

- Bernoulli: $p(x) = \alpha^{x}(1-\alpha)^{1-x} \qquad \alpha \in [0,1] \quad x \in \{0,1\}$
- Log-Likelihood (IID): $\sum_{i=1}^{N} \log p(x_i \mid \alpha) = \sum_{i=1}^{N} \log \alpha^{x_i}(1-\alpha)^{1-x_i}$
- Gradient=0:

$$\frac{\partial}{\partial \alpha}\sum_{i=1}^{N} \log \alpha^{x_i}(1-\alpha)^{1-x_i} = 0$$

$$\frac{\partial}{\partial \alpha}\sum_{i=1}^{N} x_i \log \alpha + (1-x_i)\log(1-\alpha) = 0$$

$$\frac{\partial}{\partial \alpha}\sum_{i \in class1} \log \alpha + \sum_{i \in class0} \log(1-\alpha) = 0$$

$$\sum_{i \in class1} \tfrac{1}{\alpha} - \sum_{i \in class0} \tfrac{1}{1-\alpha} = 0$$

$N_1 = \text{num}(i \in \text{class1})$

$N_2 = \text{num}(i \in \text{class2})$

$$N_1 \tfrac{1}{\alpha} - N_0 \tfrac{1}{1-\alpha} = 0$$

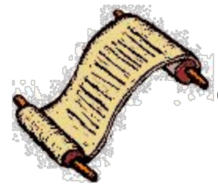
$$N_1(1-\alpha) - N_0\alpha = 0$$

$$N_1 - (N_1 + N_0)\alpha = 0$$

$$\alpha = \frac{N_1}{N_1 + N_0}$$

x=0	x=1
$\frac{N_0}{N_0 + N_1}$	$\frac{N_1}{N_0 + N_1}$

Text Modeling via Naïve Bayes



$\in \begin{cases} email \\ spam \end{cases}$

- Naïve Bayes: the simplest model of text
- There are about 50,000 words in English
- Each document is D=50,000 dimensional binary vector $\vec{x}_i$
- Each dimension is a word, set to 1 if word in the document

Dim1: "the" = 1
Dim2: "hello" = 0
Dim3: "and" = 1
Dim4: "happy" = 1
...

⇒ Do or Do not appear
Not consider multi-appearance situation

- Naïve Bayes: assumes each word is independent

$$p(\vec{x}) = p(\vec{x}(1),\ldots,\vec{x}(D)) = \prod_{d=1}^{D} p(\vec{x}(d))$$

$$= \prod_{d=1}^{D} \vec{\alpha}(d)^{\vec{x}(d)} \left(1-\vec{\alpha}(d)\right)^{\left(1-\vec{x}(d)\right)}$$

- Each 1 dimensional alpha(d) is a Bernoulli parameter
- The whole alpha vector is multivariate Bernoulli

Text Modeling via Naïve Bayes

- Maximum likelihood: assume we have several IID vectors
- Have N documents, each a 50,000 dimension binary vector
- Each dimension is a word, set to 1 if word in the document

			$\vec{x}_1$	$\vec{x}_2$	$\vec{x}_3$	$\vec{x}_4$
Dim1:	**"the"**	**=**	**1**	**0**	**1**	**1**
Dim2:	**"hello"**	**=**	**0**	**1**	**0**	**1**
Dim3:	**"and"**	**=**	**1**	**1**	**0**	**1**
Dim4:	**"happy"**	**=**	**1**	**0**	**0**	**1**

$=0$ or 1

- Likelihood $= \prod_{i=1}^{N} p\left(\vec{x}_i \mid \vec{\alpha}\right) = \prod_{i=1}^{N} \prod_{d=1}^{50000} \vec{\alpha}\left(d\right)^{\vec{x}_i(d)} \left(1 - \vec{\alpha}\left(d\right)\right)^{\left(1-\vec{x}_i(d)\right)}$

$N_d = 3$

$N = 5$

- Max likelihood solution: for each word d count number of documents it appears in divided by total N documents $\vec{\alpha}\left(d\right) = \frac{N_d}{N}$

in 3 documents, the word appears. Total number of doc. in class 1 is 5.

$\alpha_+(d) = \frac{N_{d+}}{N_+}$; $\alpha_-(d) = \frac{N_{d-}}{N_-}$

- To classify a new document x, build two models α_{+1} α_{-1} & compare $prediction = \arg\max_{y \in \{\pm 1\}} p\left(\vec{x} \mid \vec{\alpha}_y\right)$

provide $x_i(d)$

① assume $\arg\max_{y \in \{\pm 1\}} \left\{ \prod_{i=1}^{N} p(\vec{x}_i \mid \alpha_+) / \prod_{i=1}^{N} p(\vec{x}_i \mid \alpha_-) \right\}$

Categorical Probability Models

- Categorical: a distribution over a single multi-category event

1	2	3	4	5	6
$\vec{\alpha}(1)$	$\vec{\alpha}(2)$	$\vec{\alpha}(3)$	$\vec{\alpha}(4)$	$\vec{\alpha}(5)$	$\vec{\alpha}(6)$

$$p(x) = \prod_{m=1}^{M} \vec{\alpha}(m)^{\vec{x}(m)} \qquad \sum_m \vec{\alpha}(m) = 1 \qquad \vec{x} \in \mathbb{B}^M \; ; \sum_m \vec{x}(m) = 1$$

- Encode events as binary indicator vectors

$\vec{x}(1)$	$\vec{x}(2)$	$\vec{x}(3)$	$\vec{x}(4)$	$\vec{x}(5)$	$\vec{x}(6)$

- Related to the more general *multinomial* distribution
- Find α using Maximum Likelihood (with IID assumption):

$$\sum_{i=1}^{N} \log p(\vec{x}_i \mid \vec{\alpha}) = \sum_{i=1}^{N} \log \prod_{m=1}^{M} \vec{\alpha}(m)^{\vec{x}_i(m)} = \sum_{i=1}^{N} \sum_{m=1}^{M} \vec{x}_i(m) \log(\vec{\alpha}(m))$$

- Can't just take gradient over α, use sum= 1 constraint:
- Insert constraint using Lagrange multipliers

$$\frac{\partial}{\partial \alpha_q} \sum_{i=1}^{N} \sum_{m=1}^{M} \vec{x}_i(m) \log(\vec{\alpha}(m)) - \lambda\left(\sum_{m=1}^{M} \vec{\alpha}(m) - 1\right) = 0$$

$$\sum_{i=1}^{N} \left(\vec{x}_i(q) \frac{1}{\vec{\alpha}(q)} \right) - \lambda = 0 \quad \Rightarrow \quad \vec{\alpha}(q) = \tfrac{1}{\lambda} \sum_{i=1}^{N} \vec{x}_i(q)$$

Categorical Maximum Likelihood

- Taking the gradient with Lagrangian gives this formula for each q: $\vec{\alpha}(q) = \frac{1}{\lambda}\sum_{i=1}^{N}\vec{x}_i(q)$
- Recall the constraint: $\sum_m \vec{\alpha}(m) - 1 = 0$
- Plug in α's solution: $\sum_m \frac{1}{\lambda}\sum_{i=1}^{N}\vec{x}_i(m) - 1 = 0$
- Gives the lambda: $\lambda = \sum_m \sum_{i=1}^{N}\vec{x}_i(m)$
- Final answer:

$$\vec{\alpha}(q) = \frac{\sum_{i=1}^{N}\vec{x}_i(q)}{\sum_m \sum_{i=1}^{N}\vec{x}_i(m)} = \frac{N_q}{N}$$

- Example: Rolling dice
  1,6,2,6,3,6,4,6,5,6

x=1	x=2	x=3	x=4	x=5	x=6
0.1	0.1	0.1	0.1	0.1	0.5

$\boxed{\text{Class}}$

1 6 2 6 3 6 4 6 5 6

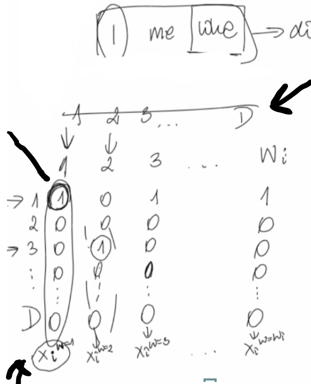
$\alpha(1) =$

$$\alpha(q) = \frac{\sum_{i=1}^{N}\vec{x}_i(q)}{\sum_m\sum_{i=1}^{N}\vec{x}_i(m)} = \frac{N_q}{N} = \boxed{\frac{1}{10}}$$

$q = 1,2,3,4,5,6$

m	1	6	2	6	3	6	4	6	5	6
1	1	0	0	0	0	0	0	0	0	0
2	0	0	1	0	0	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0
4	0	0	0	0	0	0	1	0	0	0
5	0	0	0	0	0	0	0	0	1	0
6	0	1	0	1	0	6	0	1	0	1

① For each word w_i, Seach for the dictionary with D words.
② $\vec{X}_i^w(d)$: ith document, With word D Dictionary $\in\{0,1\}$
③



words in the documents

Dictionary

Itinomial Probability Model

- The multinomial is a categorical over *counts* of events
  Dice: 1,3,1,4,6,1,1 Word Dice: the, dog, jumped, the
- Say document i has W_i=2000 words, each an IID dice roll

Dictionary

$$\cancel{p(doc_i) = p\left(\vec{x}_i^1, \vec{x}_i^2, \ldots, \vec{x}_i^{W_i}\right) = \prod_{w=1}^{W_i} p\left(\vec{x}_i^w\right) \propto \prod_{w=1}^{W_i} \prod_{d=1}^{D} \vec{\alpha}\left(d\right)^{\vec{x}_i^w(d)}}$$

1 or 0. Binary.

- Get count of each time an event occurred

$$p(doc_i) \propto \prod_{w=1}^{W_i} \prod_{d=1}^{D} \vec{\alpha}\left(d\right)^{\vec{x}_i^w(d)} = \prod_{d=1}^{D} \vec{\alpha}\left(d\right)^{\sum_{w=1}^{W_i} \vec{x}_i^w(d)} = \prod_{d=1}^{D} \vec{\alpha}\left(d\right)^{\vec{X}_i(d)}$$

- BUT: order shouldn't matter when "counting" so multiply by # of possible choosings. Choosing X(1),…X(D) from N

$$\begin{pmatrix} W_i \\ \vec{X}_i(1), \ldots, \vec{X}_i(D) \end{pmatrix} = \frac{W_i!}{\prod_{d=1}^{D} \vec{X}_i(d)!} = \frac{\left(\sum_{d=1}^{D} \vec{X}_i(d)\right)!}{\prod_{d=1}^{D} \vec{X}_i(d)!}$$

- Multinomial: over discrete integer vectors X summing to W

$$p\left(\vec{X}_i\right) = \frac{W!}{\prod_{d=1}^{D} \vec{X}(d)!} \prod_{d=1}^{D} \vec{\alpha}\left(d\right)^{\vec{X}(d)} \quad s.t. \sum_d \vec{\alpha}\left(d\right) = 1, \vec{X} \in \mathbb{Z}_+^D, \sum_{d=1}^{D} \vec{X}\left(d\right) = W$$

Tony Jebara, Columbia University

Text Modeling via Multinomial

$\in \left\{ \begin{matrix} email \\ spam \end{matrix} \right.$

- Also known as the bag-of-words model
- Each document is 50,000 dimensional vector
- Each dimension is a word, set to # times word in doc

			X_1	X_2	X_3	X_4
Dim1:	**"the"**	**=**	**9**	**3**	**1**	**0**
Dim2:	**"hello"**	**=**	**0**	**5**	**3**	**0**
Dim3:	**"and"**	**=**	**6**	**2**	**2**	**2**
Dim4:	**"happy"**	**=**	**2**	**5**	**1**	**0**

$\alpha(d) = \frac{N_q}{q}$

$X_i(d) = 0$ or 1

$\alpha(d) =$ categorical

- Each document is a vector of multinomial counts maximum likelihood

$$p\left(doc_i\right) = p\left(\vec{X}_i\right) = \frac{\left(\sum_{d=1}^{D} \vec{X}_i(d)\right)!}{\prod_{d=1}^{D} \vec{X}_i(d)!} \prod_{d=1}^{D} \vec{\alpha}(d)^{\vec{X}_i(d)} \qquad \sum_d \vec{\alpha}(d) = 1 \quad X \in \mathbb{Z}_+^D$$

- Log-likelihood: $l(\vec{\alpha}) = \sum_{i=1}^{N} \log p\left(\vec{X}_i\right) = \sum_{i=1}^{N} \log \frac{\left(\sum_{d=1}^{D} \vec{X}_i(d)\right)!}{\prod_{d=1}^{D} \vec{X}_i(d)!} \prod_{d=1}^{D} \vec{\alpha}(d)^{\vec{X}_i(d)}$

$$= \sum_{i=1}^{N} \sum_{d=1}^{D} \vec{X}_i(d) \log \vec{\alpha}(d) + const$$

- Find α just like the multinomial maximum likelihood formula!

Text Modeling Experiments

- For text modeling (McCallum & Nigam '98)
  - Bernoulli better for small vocabulary
  - Multinomial better for large vocabulary