1/31/2023: Naive Bayes

Classification Algorithms

Discriminative
- Logistic Regression
- Softmax Regression

Directly model $p(y|x)$

e.g. logistic regression

$$p(y=1|x;w) = \sigma(w^T x)$$

Don't try to model $p(x)$

Generative
- Naive Bayes

Jointly model $p(x,y)$

$$p(x,y) = \boxed{p(y)}\,\boxed{p(x|y)}$$

$p(y)$: prior distribution over labels

$p(x|y)$: Given a label, what does a plausible x look like?

$$p(y|x) = \frac{p(y)\,p(x|y)}{\boxed{p(x)}}$$

$p(x)$: Normalizing constant

$$= \sum_{k=1}^{C} p(y=k)\,p(x|y=k)$$

Naive Bayes: assume $x \in \mathbb{R}^d$

The Naive Bayes Assumption is: $P(x|y) = \prod_{j=1}^{d} P(x_j|y)$

i.e. all the x_j's are conditionally independent given y

(we don't assume they are "independent")

$p(y=0) = p(y=1) = 0.5$

Suppose $x \in \mathbb{R}^2$, $x_1, x_2 \in \{0,1\}$

$p(x_1=1|y=0) = 0.9$, $p(x_2=1|y=0) = 0.8$ → $p(x|y=0)$

$p(x_1=1|y=1) = 0.2$, $p(x_2=1|y=1) = 0.05$ → $p(x|y=1)$

Note: x_1 and x_2 not independent (non-conditionally)

If $x_1 = 1$, $y=0$ is likely $\Rightarrow$ $x_2 = 1$ is more likely

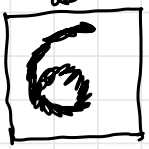
---

Common case: $x_j \in \{0,1\}$ $\forall j$

For this, we use Multivariate Bernoulli NB

many of these → distribution over $\{0,1\}$

Example 1: Black & white images

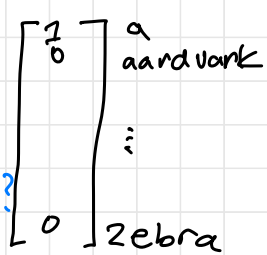


→ $28^2 = 784$-dim vector, each entry is $\{0,1\}$

Example 2: Text classification

input = document → does the word occur in document?

"bag of words"



Vocabulary V of size $|V|$

Parameters of Multivariate Bernoulli NB model

- $P(y)$: Distribution over C classes $\Rightarrow$ $\pi \in \mathbb{R}^C$ where $p(y=k) = \pi_k$

parameter #1

- $P(x_j \mid y=k)\ \forall j \in \{1, \dots, d\}$, $k \in \{1, \dots, C\}$ → Each one is a Bernoulli

So we have parameter $\tau \in \mathbb{R}^{d \times C}$ where $p(x_j = 1 \mid y=k) = \tau_{jk}$

parameter #2

How to choose π & τ? Apply MLE

$$\underbrace{LL(\pi, \tau)}_{\text{log-likelihood}} = \sum_{i=1}^{n} \log P(x^{(i)}, y^{(i)}; \pi, \tau)$$

General form for generative classifier

$$= \sum_{i=1}^{n} \log P(y^{(i)}; \pi) + \log P(x^{(i)} \mid y^{(i)}; \tau)$$

$$\sum_{i=1}^{n} \log P(y^{(i)}; \pi)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{C} \mathbb{1}[y^{(i)} = k] \log \underbrace{P(y=k; \pi)}_{=\pi_k}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{C} \mathbb{1}[y^{(i)} = k] \log \pi_k$$

Let $\text{count}(y=k)$ means $\sum_{i=1}^{n} \mathbb{1}[y^{(i)} = k]$

$$= \sum_{k=1}^{C} \text{count}(y=k) \log \pi_k$$

If $C=2$:

$$= \text{count}(y=1) \log \pi_1 + [n - \text{count}(y=1)] \cdot \log(1 - \pi_1)$$

From HW0: maximized when $\pi_1 = \dfrac{\text{count}(y=1)}{n}$

Even when $C > 2$, MLE estimate for π is

$$\pi_k = \frac{\text{Count}(y=k)}{n}$$

what about τ?

maximize $\sum_{i=1}^{n} \log P(x^{(i)} \mid y^{(i)}; \tau)$

By similar derivation, to maximize this, set

$$\tau_{jk} = \frac{\text{Count}(x_j=1, y=k)}{\text{Count}(y=k)}$$

$= P(x_j=1 \mid y=k; \tau)$

Don't actually use!!

y	x_1	x_2
1	0	1
2	1	1
3	1	0
2	0	0
2	0	1
1	1	0
2	1	0

Estimate $\tau_{11} = P(x_1=1 \mid y=1) = \frac{1}{3}$

$\tau_{21} = P(x_2=1 \mid y=1) = \frac{2}{3}$

what happens when some counts are zero?

Test classification

- "giraffe" never occurred when $y=1$
- "choir" never occurred when $y=2$

What if a document has "giraffe" and "choir"?

$P(x \mid y=1) = 0$ b/c giraffe is "impossible"

$P(x \mid y=2) = 0$ b/c choir is "impossible"

Assuming zero probability for possible events is bad

Solution: Laplace Smoothing ("pseudocounts")

↓

Pretend we've seen every (feature, label) pair λ times outside of the dataset

Hyperparameter → λ

$\lambda = 1$ reasonable

Better formula for τ_{jk}:

$$\tau_{jk} = \frac{\text{count}(y=k, x_j=1) + \lambda}{\text{count}(y=k) + 2\cdot\lambda}$$

↑ once for $(y=k, x_j=1)$, once for $(y=k, x_j=0)$

If no training data then

$$\tau_{jk} = \frac{1}{2}$$

With enough training data, ignore λ's

Another variant: Multinomial NB (for text classification)

Input $x^{(i)}$ is document = list of words w/ length d_i

By Naive Bayes Assumption:

$$P(x^{(i)} \mid y^{(i)}) = \prod_{j=1}^{d_i} P(x_j^{(i)} \mid y^{(i)})$$

Note: preserves frequency info

$P(x_j^{(i)} \mid y^{(i)})$: multinomial distribution over V (vocabulary)

Additional assumption:

$P(x_j \mid y)$ is same for all j

Distribution of 1st word | y = Distribution of 27th word | y

doc = [dog dog dog ...]

$p(\text{dog} \mid y)^3$

Parameters

- $P(y)$ — same as before: π where $p(y=k) = \pi_k$
- $P(x_j \mid y=k)$ = Distribution over vocabulary V for each k

$$P_{uk} = p(x_j = u \mid y = k)$$

$u \in V$ $\quad$ $k \in \{1, \ldots, C\}$

Y	X
1	I saw a ...
0	Hello
1	~100 words

To estimate best P_{uk}, we count

$$P_{uk} = \frac{\text{count}(x_j = u, y = k) + \lambda}{\sum_{i=1}^{n} \mathbb{1}[y^{(i)} = k] d_i + |V| \cdot \lambda}$$

d_i = # words in i-th doc

Add λ for every possible word in dictionary

whole denominator = # words that go with $y=k$