Bayesian Methods: Naïve Bayes

Rishabh Iyer
University of Texas at Dallas

based on the slides of Vibhav Gogate and Nick Rouzzi

Last Time

- Parameter learning
  - Learning the parameter of a simple coin flipping model

    Dice Rolls
- Prior distributions
- Posterior distributions
- Today: more parameter learning and naïve Bayes

Maximum Likelihood Estimation (MLE)

- **Data:** Observed set of α_H heads and α_T tails
- **Hypothesis:** Coin flips follow a binomial distribution
- **Learning:** Find the "best" θ [Ind. coin flips] $\sim B(\theta)$.
- **MLE:** Choose θ to maximize the likelihood (probability of *D* given θ)

$$\theta_{MLE} = \arg\max_{\theta} p(D|\theta)$$

$p(x_1, \ldots, x_m | \theta)$

MAP Estimation

- Choosing θ to maximize the posterior distribution is called maximum a posteriori (MAP) estimation

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|D) = \frac{p(\theta)\, p(D|\theta)}{p(D)}$$

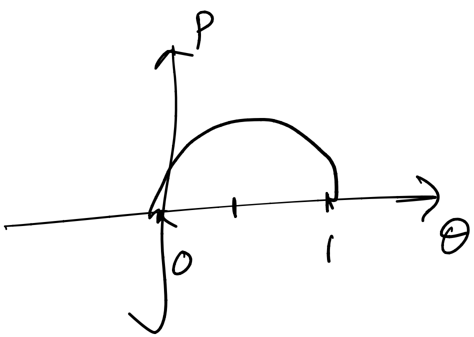
$p(D) \leftarrow$ ind of θ

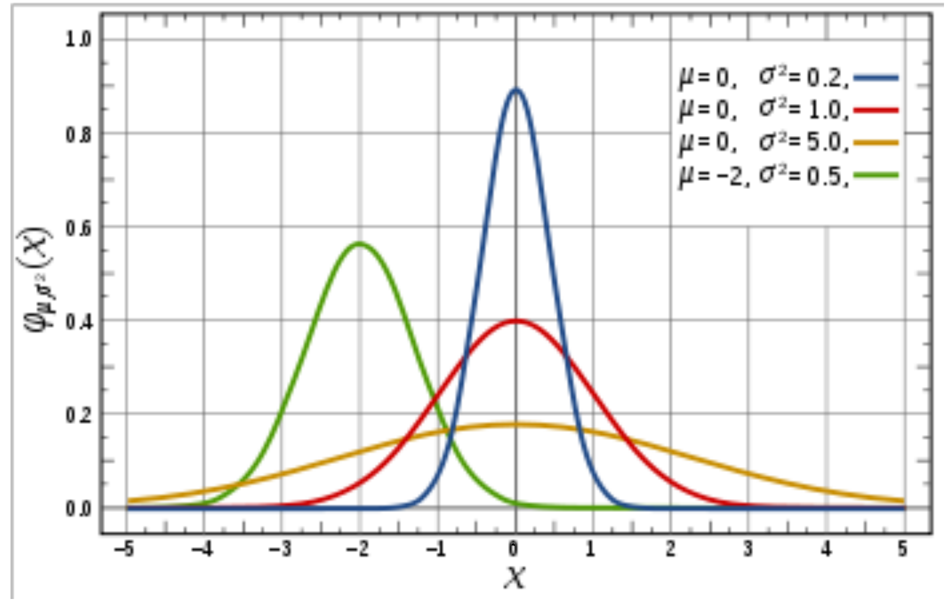
- The only difference between θ_{MLE} and θ_{MAP} is that one assumes a uniform prior (MLE) and the other allows an arbitrary prior

MLE for Gaussian Distributions

$x_1 \ldots x_M.$

- Two parameter distribution characterized by a mean and a variance





$$P(x \mid \overbrace{\underset{\uparrow\ \text{mean}}{\mu}, \underset{\uparrow\ \text{variance.}}{\sigma}}^{\theta}) = \frac{1}{\underline{\sigma\sqrt{2\pi}}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Some properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant) are Gaussian
  - $X \sim N(\mu, \sigma^2)$

    $E[Y] = aE[X] + b$
  - $Y = aX + b \Rightarrow Y \sim N(a\mu + b, a^2\sigma^2)$
- Sum of Gaussians is Gaussian
  - $X \sim N(\mu_X, \sigma_X^2),\ Y \sim N(\mu_Y, \sigma_Y^2)$
  - $Z = X + Y \Rightarrow Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$
- Easy to differentiate, as we will see soon!

Learning a Gaussian

- Collect data
  - Hopefully, i.i.d. samples
  - e.g., exam scores
- Learn parameters
  - Mean: μ
  - Variance: σ

i	Exam Score
0	85
1	95
2	100
3	12
…	…
99	89

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

MLE for Gaussian:

- Probability of N i.i.d. samples $D = x^{(1)}, \ldots, x^{(N)}$

$$p(D|\mu,\sigma) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \prod_{i=1}^{N} e^{-\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}}$$

$$\mu_{MLE}, \sigma_{MLE} = \arg\max_{\mu,\sigma} P(\mathcal{D} \mid \mu, \sigma)$$

- Log-likelihood of the data

$$\ln p(D|\mu,\sigma) = -\frac{N}{2}\ln 2\pi\sigma^2 - \sum_{i=1}^{N} \frac{\left(x^{(i)} - \mu\right)^2}{2\sigma^2}$$

MLE for the Mean of a Gaussian

$$\frac{\partial}{\partial\mu}\ln p(D|\mu,\sigma) = \frac{\partial}{\partial\mu}\left[-\frac{N}{2}\ln 2\pi\sigma^2 - \sum_{i=1}^{N}\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}\right]$$

$$= \frac{\partial}{\partial\mu}\left[-\sum_{i=1}^{N}\frac{\left(x^{(i)}-\mu\right)^2}{2\sigma^2}\right]$$

$$= \sum_{i=1}^{N}\frac{\left(x^{(i)}-\mu\right)}{\sigma^2}$$

$$= \frac{\left[N\mu - \sum_{i=1}^{N} x^{(i)}\right]}{\sigma^2} = 0$$

$$\boxed{\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}}$$

MLE for Variance

$$\frac{\partial}{\partial \sigma} \ln p(D|\mu, \sigma) = \frac{\partial}{\partial \sigma}\left[-\frac{N}{2}\ln 2\pi\sigma^2 - \sum_{i=1}^{N}\frac{\left(x^{(i)} - \mu\right)^2}{2\sigma^2}\right]$$

$$= -\frac{N}{\sigma} + \frac{\partial}{\partial \sigma}\left[-\sum_{i=1}^{N}\frac{\left(x^{(i)} - \mu\right)^2}{2\sigma^2}\right]$$

$$= -\frac{N}{\sigma} + \sum_{i=1}^{N}\frac{\left(x^{(i)} - \mu\right)^2}{\sigma^3} = 0$$

$$\boxed{\sigma_{MLE}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x^{(i)} - \mu_{MLE}\right)^2}$$

Learning Gaussian parameters

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$

$$x^{(i)} \sim N(\mu, \sigma^2)$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\left(x^{(i)} - \mu_{MLE}\right)^2$$

$$\sigma^2$$

- MLE for the variance of a Gaussian is **biased:** $E\left[\sigma^2_{MLE}\right] \neq \sigma^2_{True}$

Mean.

$$E[\mu_{MLE}] = \frac{1}{N} E\left[\sum_{i=1}^{N} x^{(i)}\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E\left[x^{(i)}\right]$$

$$= \mu$$

Unbiased.

$$E[\sigma^2_{MLE}] \neq \sigma^2$$

$$E[\sigma^2_{MLE}] = \frac{1}{N}\sum_{i=1}^{N} E\left[x^{(i)} - \mu_{MLE}\right]^2$$

$$= \frac{1}{N} E\left[\sum_{i=1}^{N} X^{(i)2} - 2N\bar{X}^2 + N\bar{X}^2\right]$$

$$\mu_{MLE} = \bar{X}$$

$$E[\sigma^2_{MLE}] = \frac{1}{N} E\left[\sum_{i=1}^{N} (x^{(i)})^2 - N\bar{x}^2\right]$$

Now, $E[x^2] = \sigma_x^2 + \mu_x^2$ & $\mu_{MLE} = \mu = \bar{x}$

$$\Rightarrow E[\sigma^2_{MLE}] = \frac{1}{N}\left[\sum_{i=1}^{N} E[x^{(i)2}] - N\bar{x}^2\right]$$

$$= E[x^2] - E[\bar{x}]^2$$

$$= \sigma^2 + \mu^2 - \sigma_{\bar{x}}^2 - \mu^2$$

$$= \sigma^2 - \sigma_{\bar{x}}^2$$

$$Var(\bar{x}) = Var\left(\frac{1}{N}\sum_{i=1}^{N} x^{(i)}\right) = \frac{1}{N^2} \cdot N\sigma^2 = \frac{1}{N}\sigma^2$$

$$\Rightarrow E[\sigma^2_{MLE}] = \frac{(N-1)\sigma^2}{N}$$

Learning Gaussian parameters

$$\mu_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$

$x^{(i)} \sim N(\mu, \sigma^2)$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\left(x^{(i)} - \mu_{MLE}\right)^2 = \frac{N-1}{N}\sigma^2_{unbiased}$$

σ^2

- MLE for the variance of a Gaussian is **biased:** $\boldsymbol{E\left[\sigma^2_{MLE}\right] \neq \sigma^2_{True}}$
  - Expected result of estimation is **not** true parameter!
  - Unbiased variance estimator

$$\sigma^2_{unbiased} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x^{(i)} - \mu_{MLE}\right)^2$$

Bayesian Categorization/Classification

- Given features $x = (x_1, \ldots, x_m)$ predict a label y
- If we had a joint distribution over x and y, given x we could find the label using MAP inference

$$\arg\max_y p(y|x_1, \ldots, x_m)$$

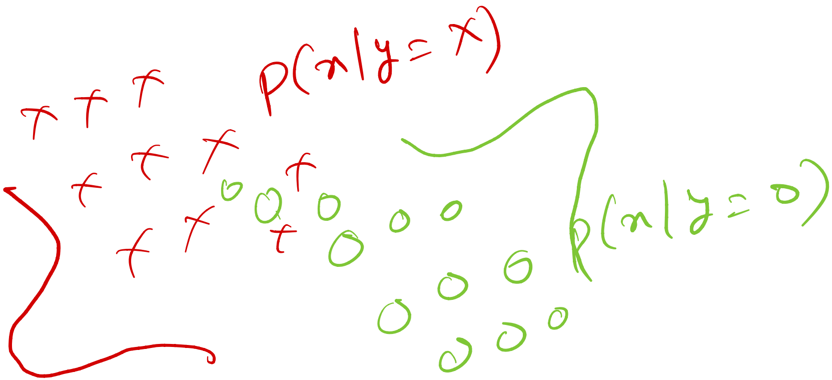
- Can compute this in exactly the same way that we did before using Bayes rule:

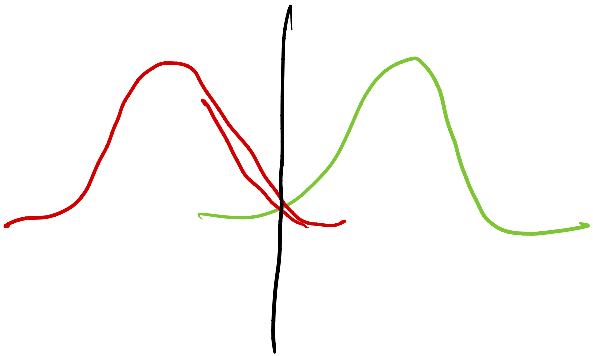
$$p(y|x_1, \ldots, x_m) = \frac{\overbrace{p(x_1, \ldots, x_m|y)p(y)}^{p(x_1, x_2, \ldots, x_m, y \mid \theta)}}{p(x_1, \ldots, x_m)}$$

① Inference.
Given x, θ
obtain y.

② Training.
Given $(x_1, y_1) \ldots (x_n, y_n)$
Obtain θ.

Modeling Distributions / Training





① $p(x_1, x_2, \ldots, x_m | y, \theta)$

② $p(y|\theta)$

Joint Dist: $p(x_1, x_2, \ldots x_m, y \mid \theta)$

Training: $[(x^{(1)}, y^{(1)}), \ldots (x^{(M)}, y^{(M)})]$

$$\arg\max_{\theta} \prod_{i=1}^{M} p(x^{(i)}, y^{(i)} \mid \theta) \quad \ldots \text{Lik}$$

$$\Rightarrow \arg\max_{\theta} \sum_{i=1}^{M} \log p(x^{(i)}, y^{(i)} \mid \theta) \quad \ldots LL$$

$$\Rightarrow \arg\max_{\theta} \sum_{i=1}^{M} \log p(x^{(i)} \mid y^{(i)}, \theta)\, p(y^{(i)} \mid \theta)$$

Inference: $\arg\max_{y} p(y \mid x_1, \ldots, x_m, \theta)$

$$\frac{p(x_1, \ldots, x_m \mid y, \theta)\, p(y \mid \theta)}{p(x_1, \ldots x_m \mid \theta)}$$

Article Classification

- Given a collection of news articles labeled by topic goal is, given an unseen news article, to predict topic
  - One possible feature vector:
    - One feature for each word in the document, in order
      - x_i corresponds to the i^{th} word
      - x_i can take a different value for each word in the dictionary

$\{(x^{(i)}, y^{(i)})\}_{i=1}^{M}$

$x^{(i)} \leftarrow$ Feature vec / Article.

$y^{(i)} \leftarrow$ TOPIC

Text Classification

(51, 100, 125, ...

...)

Article from rec.sport.hockey

```
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinic
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most
obvious candidate for pleasant surprise is Alex
Zhitnik. He came highly touted as a defensive
defenseman, but he's clearly much more than that.
Great skater and hard shot (though wish he were
more accurate). In fact, he pretty much allowed
the Kings to trade away that huge defensive
liability Paul Coffey. Kelly Hrudey is only the
biggest disappointment if you thought he was any
good to begin with. But, at best, he's only a
mediocre goaltender. A better choice would be
Tomas Sandstrom, though not through any fault of
his own, but because some thugs in Toronto decided
```

Text Classification

$x_1 \to x_2 \cdots$

- $x_1, x_2, \ldots$ is sequence of words in document
- The set of all possible features, and hence $p(y|x)$, is huge
  - Article at least 1000 words, $x = (x_1, \ldots, x_{1000})$
  - x_i represents i^{th} word in document
    - Can be any word in the dictionary – at least 10,000 words
  - $10{,}000^{1000} = 10^{4000}$ possible values
  - Atoms in Universe: $\sim 10^{80}$

Bag of Words Model

- Typically assume position in document doesn't matter

$$p(X_i = "the" | Y = y) = p(X_k = "the" | Y = y)$$

  - All positions have the same distribution
  - Ignores the order of words
  - Sounds like a bad assumption, but often works well!
- Features
  - Set of all possible words and their corresponding frequencies (number of times it occurs in the document)

Bag of Words

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

Need to Simplify Somehow

- Even with the bag of words assumption, there are too many possible outcomes
  - Too many probabilities m

$$p(\underbrace{x_1, \dots, x_m}|y)$$

- Can we assume some are the same?

$$p(x_1, x_2|y) = p(x_1|y)\, p(x_2|y)$$

  - This is a conditional independence assumption

$$p(x_1, \dots, x_m | y) = p(x_1|y)\, p(x_2|y) \dots p(x_m|y)$$

Conditional Independence

- X is **conditionally independent** of Y given Z, if the probability distribution for X is independent of the value of Y, given the value of Z

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

- Equivalent to

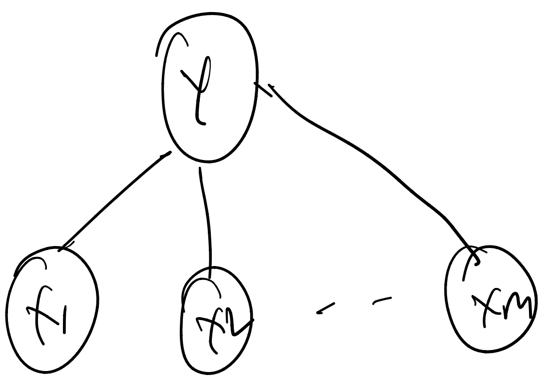
$$p(X,Y|Z) = p(X|Z)P(Y|Z)$$

$X \perp\!\!\!\perp Y$

if $p(X,Y) = p(X)p(Y)$

$X \perp\!\!\!\perp Y \mid Z$

if $p(X,Y|Z) = p(X|Z)p(Y|Z)$



$$p(x_1, x_2, \ldots, x_m \mid y) = \prod_{i=1}^{m} p(x_i \mid y)$$

Naïve Bayes

- Naïve Bayes assumption

$x_i \in [0, \dots D-1]$

  - Features are independent given class label

Size of $p(x_1, \dots x_m | y)$

$$p(x_1, x_2 | y) = p(x_1|y)\, p(x_2|y)$$

\# unique values of $(x_1, \dots, x_m)$ & y.

  - More generally

$$\boxed{\underbrace{p(x_1, \dots, x_m | y)} = \prod_{i=1}^{m} \underbrace{p(x_i | y)}}$$

size $= D^m \times |y|$

$p(x_i|y)\ \forall i$

$D \times |y| \times m$

- How many parameters now?
  - Suppose x is composed of d binary features

$p(x_1, \dots x_m | y)$ $\quad |y| = K$

$x_i \in [0, 1, \dots, D-1] \leftarrow D.$ $\quad y \in (0, \dots, K-1)$

$[x_1, x_2, \dots x_m, y]$

$0, 0, \dots 0, 0$

$0, \dots 1, 0$

$\vdots$

$p(x_1, x_2, \dots x_m | y)$

$p(x_1 = 0, x_2 = 0, \dots x_m = 0 | y)$

$\vdots$

$\rightarrow O(D^m K)$

$D^m x_1$

NBAss^m : $\dfrac{p(x_i | y)}{p(y)} \leftarrow$ For each i,

$\dfrac{Dmk}{K}$ $\dfrac{p(x_i|y)}{p(y)}$

Naïve Bayes

- Naïve Bayes assumption
  - Features are independent given class label

$$p(x_1, x_2|y) = p(x_1|y)\, p(x_2|y)$$

$p(x_i = 0|y)$

$p(x_i = 1|y)$

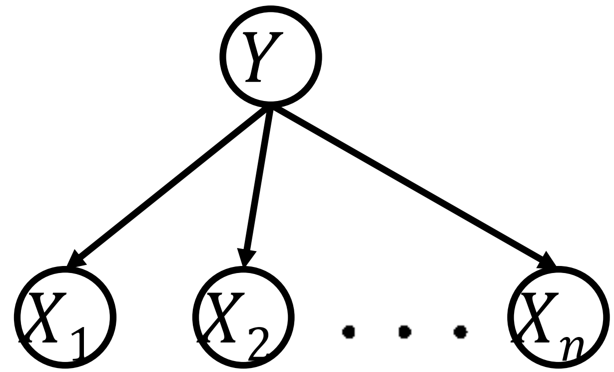
  - More generally

$$p(x_1, \ldots, x_m|y) = \prod_{i=1}^{m} p(x_i|y)$$

taking values $\in \{0, \ldots, D-1\}$

- How many parameters now?
  - Suppose x composed of $\cancel{d}$ m ~~binary~~ features $\Rightarrow O(\cancel{d \cdot L})$ $mDK)$
    where $\cancel{L}$ K is the number of class labels

The Naïve Bayes Classifier



- **Given**
  - Prior $p(y)$
  - m conditionally independent features X given the class Y
  - For each X_i, we have likelihood $P(X_i|Y)$

$\forall$ feature i

$p(x_i|y)$

$p(y)$

- Classify via

$$y^* = h_{NB}(x) = \arg\max_y \overbrace{p(y)p(x_1, \dots, x_m|y)}^{p(x_1, x_2, \dots, x_m, y)}$$

$$= \arg\max_y p(y) \prod_i^m p(x_i|y)$$

Inference

$$\arg\max_{y} \; p(y \mid x_1, x_2, \ldots, x_m)$$

$$\arg\max_{y} \; \frac{p(y, x_1, x_2, \ldots, x_m)}{p(x)}$$

$$\arg\max_{y} \; p(x_1, x_2, \ldots, x_m \mid y)\, p(y) \quad \ldots \; p(x) \text{ is const}$$

$$\arg\max_{y} \; p(y) \prod_{i=1}^{M} p(x_i \mid y)$$

MLE for the Parameters of NB

- Given dataset, count occurrences for all pairs
  - $Count(X_i = x_i, Y = y)$ is the number of samples in which $X_i = x_i$ and $Y = y$

- MLE for discrete NB

$$\theta_y = p(Y = y) = \frac{Count(Y = y)}{\sum_{y'} Count(Y = y')} \rightarrow |y| \text{ values.}$$

$$\theta_{x_i, y} = p(X_i = x_i | Y = y) = \frac{Count(X_i = x_i, Y = y)}{\sum_{x_i'} Count\,(X_i = x_i', Y = y)} \rightarrow D \text{ values}$$

See this link for more insights: http://www.datasciencecourse.org/notes/mle/

MLE For NB Classifiers

$M = \#$ train

$m = \#$ features

$$\{(x^{(1)}, y^{(1)}), \ldots (x^{(M)}, y^{(M)})\}$$

$$x^{(i)} = \{x_1^{(i)}, \ldots, x_j^{(i)}, \ldots x_m^{(i)}\}$$

$$p(D|\theta) = \prod_{i=1}^{M} p(x^{(i)}, y^{(i)} | \theta)$$

$$= \prod_{i=1}^{M} p(y^{(i)})\, p(x_1^{(i)}, x_2^{(i)}, \ldots x_m^{(i)} | y^{(i)}, \theta)$$

$$= \prod_{i=1}^{M} p(y^{(i)}) \prod_{j=1}^{m} p(x_j^{(i)} | y^{(i)})$$

$$\log p(D|\underline{\theta}) = \sum_{i=1}^{M} \left[\log p(\underline{y^{(i)}}) + \sum_{j=1}^{m} \log p(x_j^{(i)} | y^{(i)})\right]$$

$K \quad \phi_y = p(y), \quad y \in Y \quad$ e.g $\quad Y = [0, \dots K-1]$

$DKm \quad \underline{\phi_{x_i y}} = p(x_i | y), \quad y \in Y, \; x_i \in [0, \dots D-1]$

$$\log p(D|\phi) = \underbrace{\sum_{i=1}^{m} \left[\log p(y^{(i)}) + \sum_{j=1}^{m} \log p(x_j^{(i)} | y^{(i)}) \right]}$$

$C_y = \#$ [Train in D] with class "y"

$$= \sum_{y=0}^{K} C_y \log \phi_y + \sum_{i=0}^{D} \sum_{y=0}^{K} C_{x_i y} \log \phi_{x_i y}$$

$$\text{s.t} \quad \sum_{y} \phi_y = 1, \quad \sum_{i=0}^{D} \phi_{x_i y} = 1$$

$$\max_{\phi} L(\phi) \quad \text{s.t} \quad \sum_{y} \phi_y = 1, \quad \sum_{j=0}^{n} \phi_{x_i y} = 1$$

$$\{\text{"S"}, \text{"NS"}, \text{"S"}, \text{"NS"}, \text{"S"}\}$$

$$\sum_{i=1}^{5} \log p(y^{(i)})$$

$$\log p(y = \text{"S"}) + \log p(y = \text{"NS"}) + \log p(y = \text{"S"}) + \log p(y = \text{"NS"}) + \log (p = \text{"S"})$$

$$3 \log p(y = \text{"S"}) + 2 \log p(y = \text{"NS"})$$

$$\max_{\phi_y, \phi_{x|y}} \sum_y c_y \log \phi_y + \sum_i \sum_y c_{x_i|y} \log \phi_{x_i|y}$$

$$\text{s.t.} \quad \sum_y \phi_y = 1, \quad \sum_{x_i} \phi_{x_i|y} = 1$$

$$L(\phi, \lambda, M) = \sum_y c_y \log \phi_y - \lambda\left(\sum_y \phi_y - 1\right) + \cdots$$

$$\frac{\partial L}{\partial \phi} = 0 \Rightarrow \frac{c_y}{\phi_y} - \lambda = 0$$

$$\Rightarrow \phi_y = \frac{c_y}{\lambda}; \quad \text{since } \sum_y \phi_y = 1$$

$$\Rightarrow \lambda = \sum_{y'} c_{y'}$$

So,

$$\phi_y = \frac{c_y}{\sum_y c_y}$$

Similarly,
$$\phi_{x_i, y} = \frac{c_{x_i, y}}{\sum_{i=0}^{D-1} c_{x_i, y}}$$

$$p(x_1, x_2, \ldots, x_k, y) \propto p(y) \prod_{i=1}^{m} p(x_i \mid y)$$

$$\prod_{i=1}^{m} p(x^{(i)}, y \mid \phi) \propto \prod_{i=1}^{m} p(y^{(i)}) \prod_{j=1}^{d} p(x_j^{(i)} \mid y)$$

$$\propto \prod_{y=1}^{K} \phi_y^{c_y} \prod_{j=1}^{d} \prod_{x_j} \underbrace{p(x_j \mid y)}_{\phi_{x_j \mid y}}^{c(x_j, y)}$$

$$p(\phi_y) \propto \phi_y^{\alpha_y} \quad \alpha_{x_j \mid y}$$

$$p(\phi_{x_j \mid y}) \propto \phi_{x_j \mid y}$$

Explanation

$$\underbrace{\sum_{i=1}^{M} \underbrace{P(y^{(i)})}}$$

$y^{(i)} \in [0, 1]$

$P(y=0) = \phi_0$

$P(y=1) = \phi_1$

$\{(x^{(1)}, y^{(1)}), \ldots (x^{(M)}, y^{(m)})\}$

$M = 100, \quad C_0 = 60, \quad C_1 = 40$

$= \#\, \text{Ex}[y=0]\, \phi_0 + \#\, \text{Ex}(y=1)\, \phi_1$

$= 60\, \phi_0 + 40 \phi_1$

Naïve Bayes Calculations

	X_1	X_2	X_3	X_4	
Day	**Outlook**	**Temperature**	**Humidity**	**Wind**	**PlayTennis**
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$p(y = \text{"yes"}) = 9/14$, $p(y = \text{"No"}) = 5/14$.

$p(X_1 = \text{"Sunny"} \mid y = \text{"yes"}) = 2/9$

$p(X_1 = \text{"overcast"} \mid y = \text{"yes"}) = 4/9$ Training

$p(X_1 = \text{"Rain"} \mid y = \text{"yes"}) = 3/9$

⋮

Testing

$X_1 = \text{"Sunny"}$

$X_2 = \text{"Mild"}$

$X_3 = \text{"High"}$

$X_4 = \text{"Weak"}$

$$p(y = \text{"yes"} \mid X_1, X_2, X_3, X_4)$$

$$\propto p(y = \text{"yes"}, X_1, X_2, X_3, X_4)$$

$$p(X_1, X_2, X_3, X_4 \mid y = \text{"yes"})\, p(y = \text{"yes"})$$

$$= p(X_1, X_2, X_3, X_4 \mid y)\, p(y)$$

$$= p(X_1 \mid y)\, p(X_2 \mid y)\, p(X_3 \mid y)\, p(X_4 \mid y)\, p(y)$$

Subtleties of NB Classifier: #1

- Usually, features are not conditionally independent:

$$p(x_1, \dots, x_m | y) \neq \prod_{i=1}^{m} p(x_i | y)$$

- The naïve Bayes assumption is often violated, yet it performs surprisingly well in many cases

- Plausible reason: Only need the probability of the correct class to be the largest!
  - Example: binary classification; just need to figure out the correct side of 0.5 and not the actual probability (0.51 is the same as 0.99).

Subtleties of NB Classifier

- What if you never see a training instance $(X_1 = a, Y = b)$
  - Example: you did not see the word "Nigerian" in spam
    - Then $p(X_1 = a | Y = b) = 0$
    - Thus no matter what values $X_2, \cdots, X_m$ take

$$P(X_1 = \mathrm{a}, X_2 = x_2, \cdots, X_m = x_m | Y = b) = 0$$

    - Why?

Subtleties of NB Classifier

- To fix this, use a prior!
  - Already saw how to do this in the coin-flipping example using the Beta distribution
  - For NB over discrete spaces, can use the Dirichlet prior
  - The Dirichlet distribution is a distribution over $z_1, \ldots, z_k \in (0,1)$ such that $z_1 + \cdots + z_k = 1$ characterized by k parameters $\alpha_1, \ldots, \alpha_k$

$p(y \mid \alpha_1 \ldots \alpha_k)$

$$f(z_1, \ldots, z_k; \alpha_1, \ldots, \alpha_k) \propto \prod_{i=1}^{k} z_i^{\alpha_i - 1}$$

- Called smoothing, what are the MLE estimates under these kinds of priors?

Subtleties of NB Classifier

- To fix this, use a prior!
  - Already saw how to do this in the coin-flipping example using the Beta distribution
  - For NB over discrete spaces, can use the Dirichlet prior
  - The Dirichlet distribution is a distribution over $z_1, \ldots, z_k \in (0,1)$ such that $z_1 + \cdots + z_k = 1$ characterized by k parameters $\alpha_1, \ldots, \alpha_k$

$$f(z_1, \ldots, z_k; \alpha_1, \ldots, \alpha_k) \propto \prod_{i=1}^{k} z_i^{\alpha_i - 1}$$

  - Called smoothing, what are the MLE estimates under these kinds of priors?

Subtleties of NB Classifier

- To fix this, use a prior!
  - Already saw how to do this in the coin-flipping example using the Beta distribution
  - For NB over discrete spaces, can use the Dirichlet prior
  - The Dirichlet distribution is a distribution over $z_1, \ldots, z_k \in (0,1)$ such that $z_1 + \cdots + z_k = 1$ characterized by k parameters $\alpha_1, \ldots, \alpha_k$

$$f(z_1, \ldots, z_k; \alpha_1, \ldots, \alpha_k) \propto \prod_{i=1}^{k} z_i^{\alpha_i - 1}$$

    - Called smoothing, what are the MLE estimates under these kinds of priors?

$$\phi_y = \frac{C_y}{M \text{ (\# train)}}$$

$$\phi_y = \frac{C_y + \alpha_y}{M + \alpha_{y.}}$$

if $= 0$

$$\underline{\phi_{x_i|y}} = \frac{C_{x_i|y} + \alpha_y}{C_y + \alpha_{y.}}$$

without prior

$$\phi_{x_i,y} = \frac{C_{x_i y}}{C_y}$$

Train data with
x_i = overcast, y = Tennis

$$C_{x_i y} = 0$$

$$\phi_{x_i y} = 0$$

with prior

$$\phi_{x_i,y} = \frac{C_{x_i y} + \alpha_y}{C_y + \alpha_y}$$

even if $C_{x_i y} = 0$

$$\phi_{x_i y} = \frac{\alpha_y}{C_y + \alpha_y} \neq 0$$

Joint bet X & Y | Z = $P(X, Y | Z)$

Dist of X | Y, Z = $P(X | \underbrace{Y, Z})$

Continuous Naive Bayes

$$P(X_i = x \mid Y = y) = \frac{1}{\sigma_{iy}\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_{iy})^2}{2\sigma_{iy}^2}\right]$$

Some Assumptions :

$\sigma_{iy} = \sigma_i$

or $\sigma_{iy} = \sigma_y$

or $\sigma_{iy} = \sigma$

j^{th} train ex

$\downarrow$

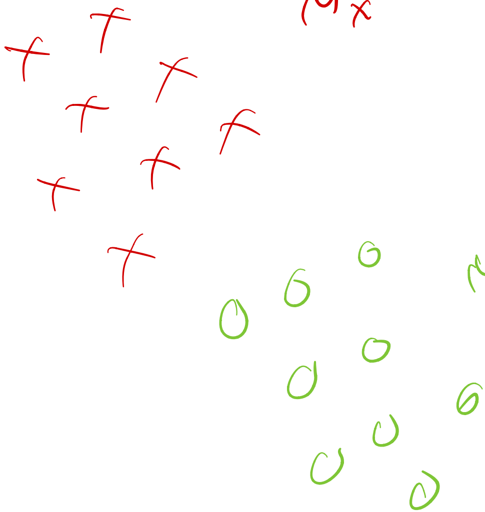
MLE :

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

$\uparrow$

$\delta(x) = 1$ if x is true.

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



$$\mu_x = \frac{\sum x_x}{\# x}$$

$$p(y = x) = \frac{\# x\text{'s}}{\# \text{points}}$$

$$\mu_o = \frac{\sum x_o}{\# o}$$

$$p(x|y), \; p(y)$$

Given x_t $\quad \arg\max_y p(y|x_t) = \arg\max_y \underbrace{p(x_t|y)} p(y)$

$$\arg\max_y p(x_{1t}|y) \dots p(x_{mt}|y)\, p(y)$$

When is Cont NB useful?

when X are Cont features

e.g Images

{predict if the brain activity corresponds to {person, Animal} using MRIs}