Naive Bayes

Dripta Mj
Department of Mathematics
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, India

Machine Learning
CS230

Sem 3, 2018-19

Introduction

- A Bayes rule based simple classification method.
- Naive because features are not usually conditionally independent.
- Popular in the field of natural language processing.
- Example: Classify an email as spam or not spam.

Example

- Want to classify an email as spam or not-spam.
- Outputs: $y \in \{c_1, c_2\}$, where c_1 indicates email is not-spam and c_2 spam.
- Training dataset: N emails (say 10000).
- Suppose we have a vocabulary comprising 5000 words.
- Dimension D of input $\mathbf{x}$ = Size of the vocabulary (=5000).
- Each email is described in terms of the vocabulary. For example, the inputs of the nth email can be expressed as

$$\mathbf{x}^{(n)} = \left[x_1^{(n)}, x_2^{(n)}, ..., x_D^{(n)}\right]^{\mathrm{T}}$$

where

$$x_j^{(n)} = \begin{cases} 1 & \text{if the } j\text{th word of vocabulary exists in the } n\text{th email} \\ 0 & \text{otherwise} \end{cases}$$

Example

- Suppose vocabulary $\mathcal{V}$ is of the form

$$\mathcal{V} = \left\{ \begin{array}{c} \text{abandon} \\ \text{ability} \\ \cdot \\ \cdot \\ \text{magic} \\ \cdot \\ \cdot \\ \text{zoo} \end{array} \right\}$$

then the feature vector

$$\mathbf{x} = \begin{bmatrix} 0 \\ 1 \\ \cdot \\ \cdot \\ 1 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

indicates that the email contains the words abiity, magic, but not abandon and zoo.

Bernoulli distribution

- Suppose x is a discrete random variable and $x \in \{0, 1\}$.
- Consider a parameter μ such that $0 \leq \mu \leq 1$.
- Let x take the value 1 with probability μ.
- Then x takes the value 0 with probability $1 - \mu$.
- The probability mass function of x can be written as

$$p(x) = \begin{cases} \mu & \text{if } x = 1 \\ 1 - \mu & \text{if } x = 0 \end{cases}$$

- The probability mass function $p(x)$ can be compactly written as

$$p(x) = \mu^x (1 - \mu)^{(1-x)}$$

- This is a Bernoulli distribution with parameter μ.

Class conditional density

- Consider a dataset comprising N data points with D features:
  - Inputs: $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}\}$
  - Outputs: $\mathbf{y} = \{y^{(1)}, y^{(2)}, ..., y^{(N)}\}$

- Suppose there are J output classes i.e. $y^{(n)} \in \{c_1, c_2, ..., c_J\}$.

  Note: In the slides $y^{(n)} \in \{1, 2, ..., J\}$ is also used to indicate class.

- Features are assumed to conditionally independent, so the model is called "naive".

- Class conditional densities as product of one-dimensional densities:

$$
\begin{aligned}
p(\mathbf{x}|y = c_j) &= p(x_1|y = c_j)\, p(x_2|y = c_j) \,..\, p(x_D|y = c_j) \\
&= \prod_{i=1}^{D} p(x_i|y = c_j)
\end{aligned}
$$

Class conditional distribution

- If we use a Bernoulli distribution to model $p(x_i|y = c_j)$ then

$$p(x_i|y = c_j) = \mu_{ij}^{x_i} \left(1 - \mu_{ij}\right)^{(1-x_i)}$$

  - Intuitively, μ_{ij} is the probability of the ith feature belonging to jth class.

- Therefore the class-conditional distribution can be written as

$$p(\mathbf{x}|y = c_j) = \prod_{i=1}^{D} \mu_{ij}^{x_i} \left(1 - \mu_{ij}\right)^{(1-x_i)}$$

Likelihood function

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X}) \propto p(\mathbf{y})p(\mathbf{X}|\mathbf{y}) &= \prod_{n=1}^{N} p(y^{(n)})p(\mathbf{x}^{(n)}|y^{(n)}) \\
&= \prod_{n=1}^{N} p(y^{(n)}) \prod_{i=1}^{D} p(x_i^{(n)}|y^{(n)}) \\
&= \prod_{n=1}^{N} p(y^{(n)}) \prod_{i=1}^{D} \mu_{iy^{(n)}}^{x_i^{(n)}} (1-\mu_{iy^{(n)}})^{(1-x_i^{(n)})} \\
&= \prod_{n=1}^{N} \pi_{y^{(n)}} \prod_{i=1}^{D} \mu_{iy^{(n)}}^{x_i^{(n)}} (1-\mu_{iy^{(n)}})^{(1-x_i^{(n)})} \\
&= L(\boldsymbol{\theta})
\end{aligned}
$$

(where $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\pi}]$ are parameters of the model)

Log likelihood

$$\log L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \left[\log \pi_{y^{(n)}} + \sum_{i=1}^{D} \log \mu_{iy^{(n)}}^{x_i^{(n)}} (1 - \mu_{iy^{(n)}})^{(1 - x_i^{(n)})} \right]$$

$$= \sum_{n=1}^{N} \left[\log \pi_{y^{(n)}} + \sum_{i=1}^{D} \log \mu_{iy^{(n)}}^{x_i^{(n)}} + \sum_{i=1}^{D} \log (1 - \mu_{iy^{(n)}})^{(1 - x_i^{(n)})} \right]$$

$$= \sum_{n=1}^{N} \left[\log \pi_{y^{(n)}} + \sum_{i=1}^{D} x_i^{(n)} \log \mu_{iy^{(n)}} + \sum_{i=1}^{D} (1 - x_i^{(n)}) \log (1 - \mu_{iy^{(n)}}) \right]$$

Maximum Likelihood Estimator

- Evaluate $\arg\max_{\boldsymbol{\theta}} \; \log L(\boldsymbol{\theta})$ subject to $\sum_j \pi_j = 1$.

- Taking derivative with respect to μ_{ij}

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \mu_{ij}} = 0$$

$$\Rightarrow \sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} \left(\frac{x_i^{(n)}}{\mu_{ij}} - \frac{(1 - x_i^{(n)})}{(1 - \mu_{ij})} \right) = 0$$

$$\Rightarrow \sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} \left[x_i^{(n)} (1 - \mu_{ij}) - (1 - x_i^{(n)}) \mu_{ij} \right] = 0$$

$$\Rightarrow \sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} \mu_{(ij)} = \sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} x_i^{(n)}$$

$$\boxed{\mu_{ij} = \frac{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} x_i^{(n)}}{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)}}}$$

Maximum Likelihood Estimator

- To derive $\boldsymbol{\pi}$, consider the Lagrangian formulation:

$$\mathcal{L} = \log L(\boldsymbol{\theta}) + \lambda(\sum_j \pi_j - 1)$$

- Taking derivative with respect to π_j

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \pi_j} + \lambda \frac{\partial \sum_j \pi_j}{\partial \pi_j} = 0$$

$$\Rightarrow \lambda = -\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} \frac{1}{\pi_j}$$

$$\pi_j = -\frac{\sum_{i=1}^{N} \mathbb{1}_{(}y^{(n)} = c_j)}{\lambda}$$

Now

$$\sum_j \pi_j = 1 \quad \Rightarrow -\frac{\sum_j \sum_{n=1}^{N} \mathbb{1}_{(}y^{(n)} = c_j)}{\lambda} = 1$$

$$\boxed{\pi_j = \frac{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)}}{N}}$$

Interpretation of parameters

$$\mu_{ij} = \frac{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} x_i^{(n)}}{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)}}$$

- μ_{ij} is the fraction of the data points assigned class c_j in which the ith feature occurs.

$$\pi_j = \frac{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)}}{N}$$

- π_j is the fraction of the training dataset assigned to the jth output class.

Prediction

- Eventually want to predict class for a new (unobserved) example with features (say) $\mathbf{x}^*$.
- Compute the posterior probability using Baye's rule. For example, the probability that $\mathbf{x}^*$ belongs to the jth class can be computed as

$$
\begin{aligned}
p(y^* = c_j|\mathbf{x}^*) &= \frac{p(\mathbf{x}^*|y^* = c_j)p(y^* = c_j)}{p(\mathbf{x}^*)} \\
&= \frac{\prod_{i=1}^{D} p(x_i^*|y^* = c_j)\pi_j}{\sum_{j=1}^{J}\prod_{i=1}^{D} p(x_i^*|y^* = c_j)\pi_j}
\end{aligned}
$$

- $\mathbf{x}^*$ is assigned to the class which has the highest posterior probability.

Email – spam or not-spam?

- Goal: For a new email with features $\mathbf{x}^*$, predict the output class y^*.
- Estimate μ_{i1}s and μ_{i2}s.
  - μ_{i1} is the fraction of the non-spam emails in which the ith word of the vocabulary occurs.
  - μ_{i2} is the fraction of the spam emails in which the ith word of the vocabulary occurs.
- Estimate π_1 and π_2.
  - π_1 is the fraction of emails which are not spam.
  - π_2 is the fraction of emails which are spam.

  Note: All the above estimations are made using the training dataset.

- Use the trained parameters to classify new emails.

GAUSSIAN MODEL

Gaussian model

- Suppose $p(\mathbf{x}|y=c_j)$ is the class likelihood and $p(y=c_j)$ is the prior of the j^{th} output class, then using Bayes rule we have

$$p(y=c_j|\mathbf{x}) = \frac{p(\mathbf{x}|y=c_j)p(y=c_j)}{\sum_{j=1}^{M} p(\mathbf{x}|y=c_j)p(y=c_j)}$$

$$\propto p(\mathbf{x}|y=c_j)p(y=c_j)$$

- Assuming conditional independence of the features given the class yields

$$p(\mathbf{x}|y=c_j) = \prod_{i=1}^{D} p(x_i|y=c_j)$$

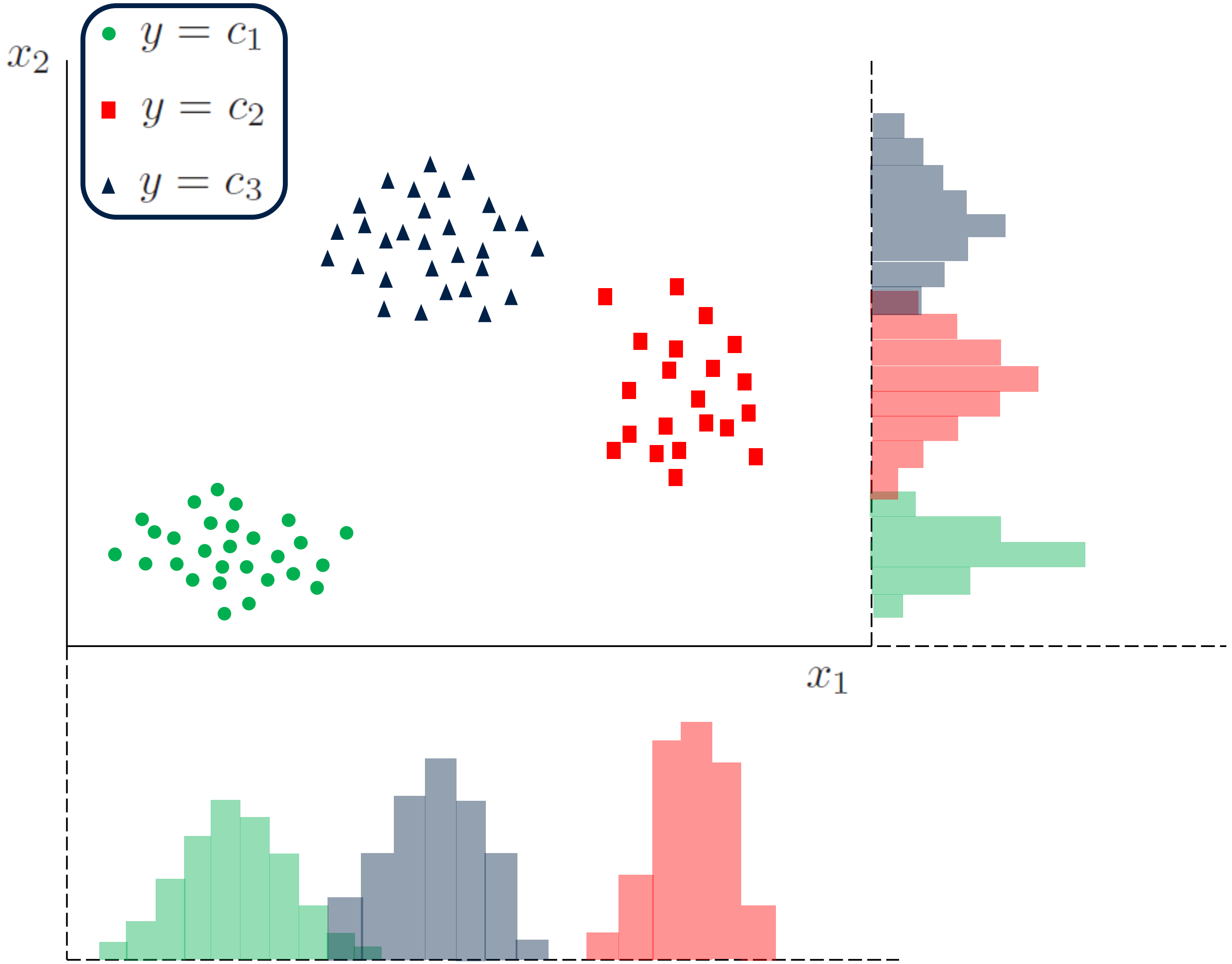
$$= \prod_{i=1}^{D} \mathcal{N}(x_i|\mu_{ij}, \sigma_{ij}^2)$$

Gaussian model

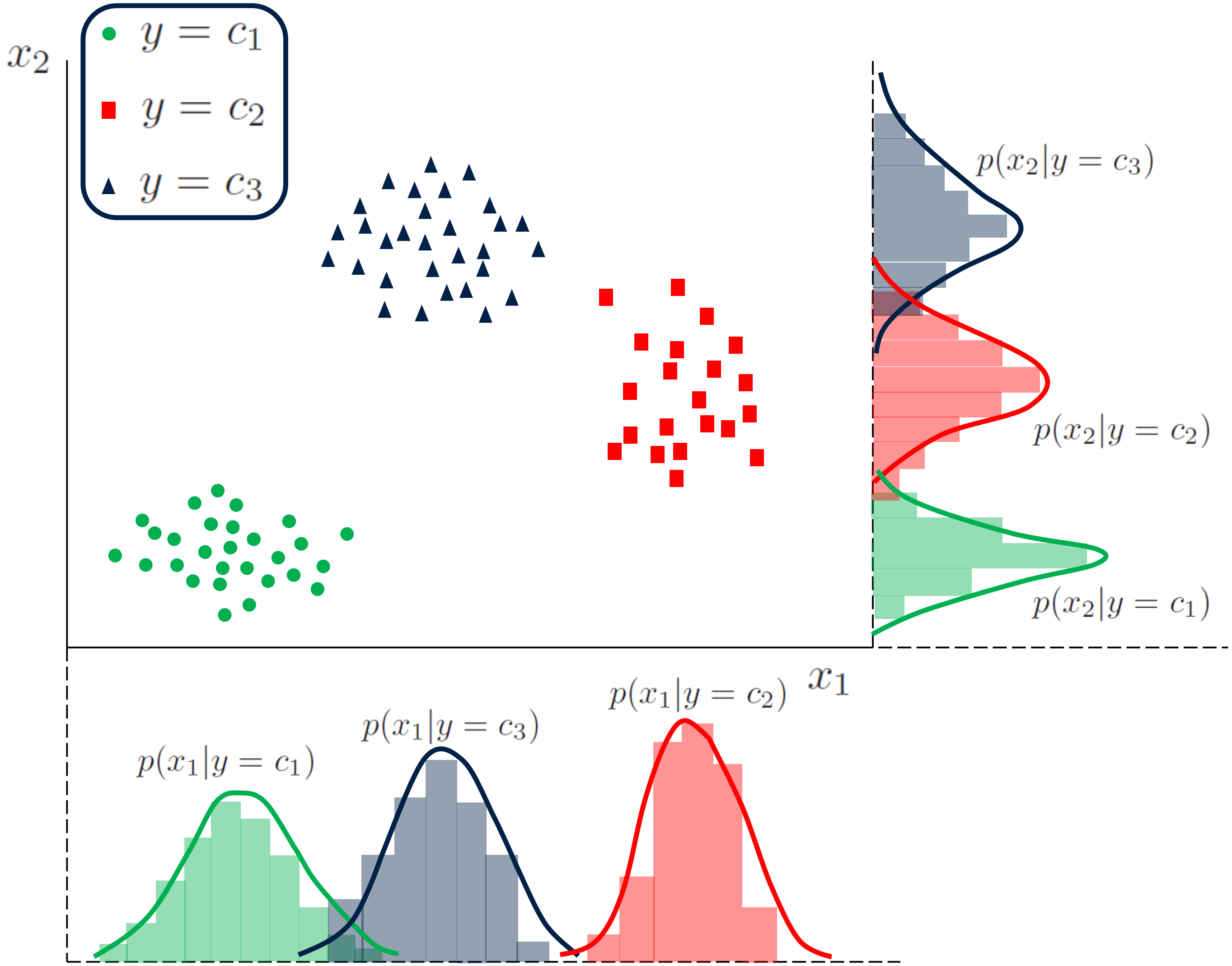
$$p(x_i|y = c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(- \frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2} \right)$$

- μ_{ij} is the mean of the ith feature belonging to class c_j.
- σ_{ij} is the variance of the ith feature belonging to class c_j.
- The model parameters are $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}]$.

Gaussian model



Gaussian model



Gaussian model – likelihood

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{X}) \propto p(\mathbf{y})p(\mathbf{X}|\mathbf{y}) &= \prod_{n=1}^{N} p(y^{(n)})p(\mathbf{x}^{(n)}|y^{(n)}) \\
&= \prod_{n=1}^{N} p(y^{(n)}) \prod_{i=1}^{D} p(x_i^{(n)}|y^{(n)}) \\
&= \prod_{n=1}^{N} p(y^{(n)}) \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi}\sigma_{iy^{(n)}}} \exp\left[-\frac{(x_i - \mu_{iy^{(n)}})^2}{2\sigma_{iy^{(n)}}^2}\right] \\
&= \prod_{n=1}^{N} \pi_{y^{(n)}} \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi}\sigma_{iy^{(n)}}} \exp\left[-\frac{(x_i - \mu_{iy^{(n)}})^2}{2\sigma_{iy^{(n)}}^2}\right] \\
&= L(\boldsymbol{\theta})
\end{aligned}
$$

(where $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\pi}]$ are parameters of the model)

Gaussian model – log likelihood

$$L(\boldsymbol{\theta}) = \prod_{n=1}^{N} \pi_{y(n)} \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi}\sigma_{iy(n)}} \exp\left[- \frac{(x_i - \mu_{iy(n)})^2}{2\sigma_{iy(n)}^2} \right]$$

- Taking logarithm on both sides

$$\log L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \left[\log \pi_{y(n)} + \sum_{i=1}^{D} \log \left(\frac{1}{\sqrt{2\pi}\sigma_{iy(n)}} \right) - \sum_{i=1}^{D} \frac{(x_i - \mu_{iy(n)})^2}{2\sigma_{iy(n)}^2} \right]$$

$$= \sum_{n=1}^{N} \left[\log \pi_{y(n)} - \sum_{i=1}^{D} \log \left(\sqrt{2\pi}\sigma_{iy(n)} \right) - \sum_{i=1}^{D} \frac{(x_i - \mu_{iy(n)})^2}{2\sigma_{iy(n)}^2} \right]$$

Gaussian model

- Evaluate $\arg\max_{\boldsymbol{\theta}} \ \log L(\boldsymbol{\theta})$ subject to $\sum_j \pi_j = 1$.

- Taking derivative with respect to μ_{ij}

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \mu_{ij}} = 0$$

$$\Rightarrow \sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} \left(\frac{(x_i^{(n)} - \mu_{ij})}{\sigma_{ij}^2} \right) = 0$$

$$\boxed{\mu_{ij} = \frac{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} x_i^{(n)}}{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)}}}$$

Gaussian model

- Taking derivative with respect to σ_{ij}

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \sigma_{ij}} = 0$$

$$\Rightarrow \sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} \left[-\frac{\sqrt{2\pi}}{\sqrt{2\pi}\sigma_{ij}} - \frac{(x_i^{(n)} - \mu_{ij})^2}{2} \left(\frac{-2}{\sigma_{ij}^3} \right) \right] = 0$$

$$\Rightarrow \sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} \left[\frac{(x_i^{(n)} - \mu_{ij})^2}{\sigma_{ij}^2} - 1 \right] = 0$$

$$\boxed{\sigma_{ij}^2 = \frac{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} (x_i^{(n)} - \mu_{ij})^2}{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)}}}$$

- π_j's can be estimated using a Lagrange formulation (as in case of Bernoulli model)

$$\boxed{\pi_j = \frac{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)}}{N}}$$

Multinomial Model

Shortcomings of Bernoulli model

- In the Bernoulli model, the feature vector (say $\mathbf{x}$) captures the presence or absence of features in the dataset.
- Such representation cannot capture the frequency of features in an example, i.e. multiple occurrence of a particular feature in the same example is not accounted.
- Multinomial model addressed this issue.
- First a look at multinomial coefficient....

Multinomial coefficient

- Number of distinct ways to permute a set of items (say a multiset of m elements).
  - Consider the word **KOLKATA**.
  - The word has 2 **K**s, 1 **O**, 1 **L**, 2 **A**s and 1 **T**.
  - The number of unique permutations of letters of the word:

$$\frac{7!}{2!\,1!\,1!\,2!\,1!}$$

Multinomial coefficient

- Suppose there are m items of D number of types.
  - Let us say there are z_i number of items of type i with $i = 1, 2, ..., D$.
  - So $z_1 + z_2 + ... + z_D = m$.
- The number of unique arrangements of the items is given by

$$\frac{m!}{z_1!\, z_2! \,....\, z_D!}$$

- This is the general form of the multinomial coefficient.
- Also, proportion of items of type i is

$$p_i = \frac{z_i}{m}$$

and therefore

$$\sum_{i=1}^{D} p_i = 1$$

Multinomial distribution

- Suppose m items are selected at random from a set comprising D types of items.
- Also suppose there are z_i number of items of kind i, such that
$$z_1 + z_2 + ... + z_D = m$$
- Let the probability of the i^{th} kind of item be p_i, and so $\sum_{i=1}^{D} p_i = 1$.
- The probability distribution of $\mathbf{z} = [z_1, z_2, ..., z_D]^{\mathrm{T}}$ is given by
$$p(\mathbf{z}) = \frac{m!}{z_1!\, z_2! \dots z_D!} (p_1)^{z_1} (p_2)^{z_2} \dots (p_D)^{z_D}$$
  - The product of the probabilities indicates the probability of having z_i number of items with feature i, with $i = 1, .., D$.
  - The multinomial coefficient indicates the number of all possible combinations with features having counts $\mathbf{z} = [z_1, z_2, .., z_D]^{\mathrm{T}}$.

Multinomial model – example

- Consider the problem of document classification: spam or not-spam.

- Suppose you are given a set of N documents which are already classified.

- The dataset can represented as $\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), ..., (\mathbf{x}^{(N)}, y^{(N)})\}$.
  - The n^{th} input data point is $\mathbf{x}^{(n)} = [x_1^{(n)}, x_2^{(n)}, .., x_D^{(n)}]^{\mathrm{T}}$.
  - $\{x_1, x_2, .., x_D\}$ are the features of the model.
  - The features correspond to a vocabulary (dictionary) of words.
  - $x_i^{(n)}$ indicates the number of times the ith word occurs in the nth email.
  - The outputs are either spam $y^{(n)} = 0$ or not-spam $y^{(n)} = 1$.

Multinomial model – likelihood function

$$p(\mathbf{y}|\mathbf{X}) \propto p(\mathbf{y})p(\mathbf{X}|\mathbf{y}) = \prod_{n=1}^{N} p(y^{(n)})p(\mathbf{x}^{(n)}|y^{(n)})$$

- The distribution $p(\mathbf{x}^{(n)}|y^{(n)})$ can be written as a multinomial distribution

$$p(\mathbf{x}^{(n)}|y^{(n)}) = \frac{m!}{x_1^{(n)}!\, x_2^{(n)}! \ldots.\, x_D^{(n)}!} \left(p(x_1|y^{(n)})\right)^{x_1^{(n)}} \left(p(x_2|y^{(n)})\right)^{x_2^{(n)}} \ldots. \left(p(x_D|y^{(n)})\right)^{x_D^{(n)}}$$

- Taking $p(x_i|y^{(n)}) = \mu_{iy^{(n)}}$, can write

$$p(\mathbf{x}^{(n)}|y^{(n)}) = \frac{m!}{x_1^{(n)}!\, x_2^{(n)}! \ldots.\, x_D^{(n)}!} \mu_{1y^{(n)}}^{x_1^{(n)}} \mu_{2y^{(n)}}^{x_2^{(n)}} \ldots. \mu_{Dy^{(n)}}^{x_D^{(n)}}$$

- Since the multinomial coefficient in front is a constant, can write

$$p(\mathbf{x}^n|y^{(n)}) \propto \prod_{i=1}^{D} \mu_{iy^{(n)}}^{x_i^{(n)}}$$

Multinomial model – likelihood function

$$p(\mathbf{y}|\mathbf{X}) \propto p(\mathbf{y})p(\mathbf{X}|\mathbf{y}) = \prod_{n=1}^{N} p(y^{(n)})p(\mathbf{x}^{(n)}|y^{(n)}) \propto \prod_{n=1}^{N} p(y^{(n)}) \prod_{i=1}^{D} \mu_{iy^{(n)}}^{x_i^{(n)}}$$

$$= \prod_{n=1}^{N} \pi_{y^{(n)}} \prod_{i=1}^{D} \mu_{iy^{(n)}}^{x_i^{(n)}}$$

$$= L(\boldsymbol{\theta})$$

- $\boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\pi}]$ are the parameters of the model.
- Log likelihood:

$$\log L(\boldsymbol{\theta}) = \sum_{n=1}^{N} \left[\log \pi_{y^{(n)}} + \sum_{i=1}^{D} \log \mu_{iy^{(n)}}^{x_i^{(n)}} \right]$$

$$= \sum_{n=1}^{N} \left[\log \pi_{y^{(n)}} + \sum_{i=1}^{D} x_i^{(n)} \log \mu_{iy^{(n)}} \right]$$

Multinomial model – MLE

- Evaluate $\arg\max_{\boldsymbol{\theta}} \; \log L(\boldsymbol{\theta})$

$$\text{subject to} \quad \sum_{i=1}^{D} \mu_{ij} = 1, \quad \forall j = 1, 2, ..., M$$

$$\sum_{j=1}^{M} \pi_j = 1$$

- Method of Lagrange multipliers:

$$\mathcal{L} = \sum_{n=1}^{N} \left[\log \pi_{y^{(n)}} + \sum_{i=1}^{D} x_i^{(n)} \log \mu_{iy^{(n)}} \right]$$

$$+\alpha_1 \Big(\sum_{i=1}^{D} \mu_{i1} - 1 \Big) + \alpha_2 \Big(\sum_{i=1}^{D} \mu_{i2} - 1 \Big) + \ldots\ldots + \alpha_M \Big(\sum_{i=1}^{D} \mu_{iM} - 1 \Big)$$

$$+\beta \Big(\sum_{j=1}^{M} \pi_j - 1 \Big)$$

Model parameters

- Taking derivative with respect to μ_{ij}:

$$\frac{\partial \mathcal{L}}{\partial \mu_{ij}} = 0$$

$$\Rightarrow \sum_{n=1}^{N} \left[\mathbb{1}_{(y^{(n)}=c_j)} \left(\frac{x_i^{(n)}}{\mu_{ij}} \right) \right] + \alpha_j = 0$$

$$\Rightarrow \alpha_j = - \sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} \frac{x_i^{(n)}}{\mu_{ij}}$$

$$\Rightarrow \mu_{ij} = - \sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} \frac{x_i^{(n)}}{\alpha_j}$$

But α_j is unknown.

Model parameters

- Substituting μ_{ij} in

$$\sum_{i=1}^{D} \mu_{ij} = 1$$

we have

$$\Rightarrow -\sum_{i=1}^{D}\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} \frac{x_i^{(n)}}{\alpha_j} = 1$$

$$\Rightarrow -\sum_{i=1}^{D}\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} x_i^{(n)} = \alpha_j$$

$$\Rightarrow -\sum_{i=1}^{D}\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} x_i^{(n)} = -\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} \frac{x_i^{(n)}}{\mu_{ij}}$$

$$\boxed{\mu_{ij} = \frac{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} x_i^{(n)}}{\sum_{i=1}^{D}\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} x_i^{(n)}}}$$

Model parameters

- π_j's can be derived in a similar way by differentiating $\mathcal{L}$ with respect to π_j, which gives

$$\boxed{\pi_j = \frac{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)}}{N}}$$

- Use the trained parameters to predict the output class of a new (unobserved) example.

Issue

- Posterior probability:

$$p(c_j|\mathbf{x}^*) \propto p(c_j)\prod_{i=1}^{D} p(x_i^*|c_j)$$
$$\propto \pi_j \prod_{i=1}^{D} \mu_{ij}^{x_i^*}$$

- Problem: If any $p(x_i^*|c_j) = 0$, then the posterior probability becomes 0.

- Example:
  - Consider the problem of classifying documents between two classes: **Science** and **Sports**.
  - Suppose the word "laser" is there in the vocabulary, but not present in the training dataset.
  - If the word "laser" occurs in a test document, then the posterior probabilities of both classes become equal to 0.

- A word not present in the training dataset of a particular document class does not imply that the word cannot occur in any document belonging to that class.

Laplace smoothing

- Derived earlier:

$$\mu_{ij} = \frac{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} x_i^{(n)}}{\sum_{i=1}^{D} \sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} x_i^{(n)}}$$

- Laplace Smoothing: Add 1 to the numerator and D (size of the vocabulary) to the denominators.

$$\mu_{ij} = \frac{\sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} x_i^{(n)} + 1}{\sum_{i=1}^{D} \sum_{n=1}^{N} \mathbb{1}_{(y^{(n)}=c_j)} x_i^{(n)} + D}$$

- $\sum_{i=1}^{D} \mu_{ij} = 1$ still holds.