Machine Learning

Lecture 5: Gaussian Discriminant Analysis, Naive Bayes and EM Algorithm

Feng Li

fli@sdu.edu.cn
https://funglee.github.io

School of Computer Science and Technology
Shandong University

Fall 2018

Probability Theory Review

- Sample space, events and probability
- Conditional probability
- Random variables and probability distributions
- Joint probability distribution
- Independence
- Conditional probability distribution
- Bayes' Theorem
- ... ...

Sample Space, Events and Probability

- A **sample space** $\mathcal{S}$ is the set of all possible outcomes of a (conceptual or physical) random experiment
- **Event** A is a subset of the sample space $\mathcal{S}$
- $P(A)$ is the **probability** that event A happens
    - It is a function that maps the event A onto the interval $[0, 1]$.
    - $P(A)$ is also called the probability measure of A
- Kolmogorov axioms
    - Non-negativity: $p(A) \geq 0$ for each event A
    - $P(\mathcal{S}) = 1$
    - σ-additivity: For disjoint events $\{A_i\}_i$ such that $A_i \bigcap A_j = \emptyset$ for $\forall i \neq j$

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Sample Space, Events and Probability (Contd')

- Some consequences
  - $P(\emptyset) = 0$
  - $P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$
  - $P(A^-) = 1 - P(A)$

Conditional Probability

- Definition of conditional probability: Fraction of worlds in which event A is true given event B is true

$$P(A \mid B) = \frac{P(A, B)}{P(B)}, \quad P(A, B) = P(A \mid B)P(B)$$

- Corollary: The chain rule

$$P(A_1, A_2, \cdots, A_k) = \prod_{k=1}^{n} P(A_k \mid A_1, A_2, \cdots, A_{k-1})$$

- Example:

$$P(A_4, A_3, A_2, A_1) = P(A_4 \mid A_3, A_2, A_1)P(A_3 \mid A_2, A_1)P(A_2 \mid A_1)P(A_1)$$

Conditional Probability (Contd.)

- Real valued **random variable** is a function of the outcome of a randomized experiment

$$X : \mathcal{S} \to R$$

- Examples: Discrete random variables ($\mathcal{S}$ is discrete)
  - $X(s) = True$ if a randomly drawn person (s) from our class ($\mathcal{S}$) is female
  - $X(s) = $ The hometown $X(s)$ of a randomly drawn person (s) from ($\mathcal{S}$)
- Examples: Continuous random variables ($\mathcal{S}$ is continuous)
  - $X(s) = r$ be the heart rate of a randomly drawn person s in our class $\mathcal{S}$

Random Variables

- Real valued **random variable** is a function of the outcome of a randomized experiment

$$X : \mathcal{S} \to R$$

- For continuous random variable X

$$P(a < X < b) = P(\{s \in \mathcal{S} : a < X(s) < b\})$$

- For discrete random variable X

$$P(X = x) = P(\{s \in \mathcal{S} : X(s) = x\})$$

Probability Distribution

- Probability distribution for **discrete** random variables
  - Suppose X is a discrete random variable

  $$X : \mathcal{S} \to \mathcal{A}$$

  - Probability mass function (PMF) of X: the probability of $X = x$

  $$p_X(x) = P(X = x)$$

  - Since $\sum_{x \in \mathcal{A}} P(X = x) = 1$, we have

  $$\sum_{x \in \mathcal{A}} p_X(x) = 1$$

Probability Distribution (Contd.)

- Probability distribution for **continuous** random variables
  - Suppose X is a continuous random variable

  $$X : \mathcal{S} \to \mathcal{A}$$

  - Probability density function (PDF) of X is a function $f_X(x)$ such that for $\forall a, b \in \mathcal{A}$ with $(a \leq b)$

  $$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

Joint Probability Distribution

- Joint probability distribution
  - Suppose both X and Y are **discrete** random variable
  - Joint probability mass function (PMF)

  $$p_{X,Y}(x, y) = P(X = x, Y = y)$$

  - Marginal probability mass function for discrete random variables

  $$p_X(x) = \sum_y P(X = x, Y = y) = \sum_y P(X = x \mid Y = y)P(Y = y)$$

  $$p_Y(y) = \sum_x P(X = x, Y = y) = \sum_x P(Y = y \mid X = x)P(Y = x)$$

  - Extension to multiple random variables $X_1, X_2, X_3, \cdots, X_n$

  $$p(x_1, x_2, \cdots, x_n) = P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$

Joint Probability Distribution (Contd.)

- Joint probability distribution
    - Suppose both X and Y are **continuous** random variable
    - Joint probability density function (PDF) $f(x, y)$

    $$P(a_1 \leq X \leq b_1, a_2 \leq Y \leq b_2) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x, y) dx dy$$

    - Marginal probability density functions

    $$
    \begin{aligned}
    f_X(x) &= \int_{-\infty}^{\infty} f(x, y) dy \text{ for } -\infty < x < \infty \\
    f_Y(x) &= \int_{-\infty}^{\infty} f(x, y) dx \text{ for } -\infty < y < \infty
    \end{aligned}
    $$

    - Extension to more than two random variables

    $$P(a_1 \leq X_1 \leq b_1, \cdots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \cdots, x_n) dx_1 \cdots dx_n$$

Independent Random Variables

- Two **discrete** random variables X and Y are **independent** if for any pair of x and y

$$p_{X,Y}(x,y) = p_X(x)p_Y(y)$$

- Two **continuous** random variables X and Y are **independent** if for any pair of x and y

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

- If the above equations do **not** hold for all (x,y), then X and Y are said to be **dependent**

Conditional Probability Distribution

- Discrete random variables X and Y
- Joint PMF $p(x, y)$
- Marginal PMF $p_X(x) = \sum_y p(x, y)$
- The **Conditional probability density function** of Y given $X = x$

$$p_{Y|X}(y \mid x) = \frac{p(x, y)}{p_X(x)}, \quad \forall y$$

or

$$p_{Y|X=x}(y) = \frac{p(x, y)}{p_X(x)}, \quad \forall y$$

- Conditional probability of $Y = y$ given $X = x$

$$P_{Y=y|X=x} = p_{Y|X}(y \mid x)$$

Conditional Probability Distribution (Contd.)

- Continuous random variables X and Y
- Joint PDF $f(x, y)$
- Marginal PDF $f_X(x) = \int_y f(x, y) dy$
- The **Conditional probability density function** of Y given $X = x$

$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)}, \quad \forall y$$

or

$$f_{Y|X=x}(y) = \frac{f(x, y)}{f_X(x)}, \quad \forall y$$

- Probability of $a \leq X \leq b$ given $Y = y$

$$P(a_1 \leq X \leq b_1 \mid Y = y) = \int_a^b f_{X|Y=y}(x) dx$$

Conditional Probability Distribution (Contd.)

- Continuous random variables X
- Discrete random variable Y
- Joint probability distribution

$$P(a \leq X \leq b, Y = y) = P(a \leq X \leq b \mid Y = y)P(Y = y)$$

where

$$P(a \leq X \leq b \mid Y = y) = \int_a^b f_{X \mid Y = y}(x)dx$$
$$P(Y = y) = p_Y(y)$$

Bayes' Theorem

- Bayes' theorem (or Bayes' rule) describes the probability of an event, based on prior knowledge of conditions that might be related to the event

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- In the Bayesian interpretation, probability measures a "degree of belief", and Bayes' theorem links the degree of belief in a proposition before and after accounting for evidence.

- For proposition A and evidence B
  - $P(A)$, the prior, is the initial degree of belief in A
  - $P(A \mid B)$, the posterior, is the degree of belief having accounted for B

- Another form:

$$P(A \mid B) = \frac{P(A \mid B)p(A)}{p(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^\neg)P(A^\neg)}$$

with $A^\neg$ being the complement of A

Bayes' Theorem (Contd')

- A: you have the flu
- B: you just coughed
- Assume: $P(A) = 0.05$, $P(B \mid A) = 0.8$, and $P(B|A^\neg) = 0.2$ ($A^\neg$ denotes of the complement of A)
- Question: $P(flue \mid cough) = P(A \mid B)$?

$$
\begin{aligned}
P(A \mid B) &= \frac{P(B \mid A)P(A)}{P(B)} \\
&= \frac{P(B \mid A)P(A)}{P(B \mid A)p(A) + P(B \mid A^\neg)P(A^\neg)} \\
&= \frac{0.8 * 0.05}{0.8 * 0.05 + 0.2 * 0.95} \\
&\approx 0.18
\end{aligned}
$$

Bayes' Theorem (Contd')

- Random variables X and Y, both of which are discrete

$$P(X = x \mid Y = y) = \frac{P(Y = y \mid X = x)P(X = x)}{P(Y = y)}$$

- Conditional PMF of X given $Y = y$

$$p_{X \mid Y = y}(x) = \frac{p_{Y \mid X = x}(y)p_X(x)}{p_Y(y)}$$

Bayes' Theorem (Contd')

- Continuous random variable X and discrete random variable Y

$$
\begin{aligned}
f_{X|Y=y}(x) &= \frac{P(Y=y \mid X=x)f_X(x)}{P(Y=y)} \\
&= \frac{p_{Y|X=x}(y)f_X(x)}{p_Y(y)}
\end{aligned}
$$

Bayes' Theorem (Contd')

- Discrete random variable X and continuous random variable Y

$$P(X = x \mid Y = y) \quad = \quad \frac{f_{Y|X=x}(y)P(X = x)}{f_Y(y)}$$

- In another form

$$p_{X|Y=y}(x) = \frac{f_{Y|X=x}(y)p(x)}{f_Y(y)}$$

Bayes' Theorem (Contd')

- X and Y are both continuous

$$f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y)f_X(x)}{f_Y(y)}$$

Prediction Based on Bayes' Theorem

- X is a random variable indicating the feature vector
- Y is a random variable indicating the label
- We perform a trial to obtain a sample x for test, and what is

$$P(Y = y \mid X = x) = p(y \mid x) \text{ ?}$$

Prediction Based on Bayes' Theorem (Contd')

- We compute $p(y \mid x)$ based on Bayes' Theorem

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)}, \ \forall y$$

- We calculate $p(x \mid y)$ for $\forall x, y$ and $p(y)$ for $\forall y$ according to the given training data

- Fortunately, we do not have to calculate $p(x)$, because

$$\arg \max_y p(y \mid x) = \arg \max_y \frac{p(x \mid y)p(y)}{p(x)} = \arg \max_y p(x \mid y)p(y)$$

Warm Up

- The world is probabilistic
  - You randomly join SDU
  - You randomly choose this class
  - You may randomly fail this class

Warm Up (Contd')

- Task: Identify if there is a cat in a given image.

Warm Up (Contd')

- Images (to be classified or to be used for training) are given randomly
  - Some of them may contain a cat
  - Some of them may not
  - Whether there is a cat is random
- An image is represented by a vector of features
- The feature vectors are random, since the images are randomly given
  - Random variable X representing the feature vector (and thus the image)
- The labels are random, since the images are randomly given
  - Random variable Y representing the label

Warm Up (Contd')

- In linear regression and logistic regression, x and y are linked through (deterministic) hypothesis function

$$y = h_\theta(x)$$

- How to model the (probabilistic) relationship between feature vector X and label Y?

$$P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)}$$

Warm Up (Contd')

- How to model the (probabilistic) relationship between feature vector X and label Y?

$$P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)}$$

  - $P(Y = y \mid X = x)$: Given an image $X = x$ (whose feature is x), what is the probability of $Y = y$ (with $y = 1$ denoting there is a cat and $y = 0$ denoting there is not)?
  - $P(X = x \mid Y = y)$: Given an image with $Y = y$ (whose label is y), what is the probability that the image has its feature vector being $X = x$?
  - $P(Y = y)$: Given a randomly picked image, what is the probability that the image contains a cat?
  - $P(X = x)$: Given a randomly picked image, what is the probability that the image has its feature vector being $X = x$?

Warm Up (Contd')

- How to model the (probabilistic) relationship between feature vector X and label Y?

$$P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)}$$

- To predict y, we have to know
    - $P(X = x \mid Y = y)$: Given an image with $Y = y$ (whose label is y), what is the probability that the image has its feature vector being $X = x$?
    - $P(Y = y)$: Given a randomly picked image, what is the probability that the image contains a cat?
    - $P(X = x)$: Given a randomly picked image, what is the probability that the image has its feature vector being $X = x$?

Warm Up (Contd')

- In our case
  - If $P(Y = 1 \mid X = x) \geq P(Y = 0 \mid X = x)$, we conclude that there is a cat
  - If $P(Y = 0 \mid X = x) \geq P(Y = 1 \mid X = x)$, we conclude there is not a cat
- How to compare $P(Y = 0 \mid X = x)$ and $P(Y = 1 \mid X = x)$

$$P(Y = 0 \mid X = x) = \frac{P(X = x \mid Y = 0)P(Y = 0)}{P(X = x)}$$

$$P(Y = 1 \mid X = x) = \frac{P(X = x \mid Y = 1)P(Y = 1)}{P(X = x)}$$

- We do not need to know $P(X = x)$

Warm Up (Contd')

- We make classification according to

$$P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y)P(Y = y)}{P(X = x)}$$

- To make classification, we have to know the following parameters

$$P(X = x \mid Y = y), \quad \forall x, y$$

$$P(Y = y), \quad \forall y$$

- We do not need to know $P(X = x), \quad \forall x$

Warm Up (Contd')

- The probability that an image labeled by y has feature vector x

$$P(X = x \mid Y = y) = p(x \mid y), \quad \forall x, y$$

- The probability that an image is labeled by y

$$P(Y = y) = p(y), \quad \forall y$$

- We compute the above parameters by learning from training data, but How?

Warm Up (Contd')

- Given a set of training data $\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1,\cdots,m}$
- The training data are sampled in an i.i.d. manner
  - The probability of the i-th training data $(x^{(i)}, y^{(i)})$

$$
\begin{aligned}
& P(X = x^{(i)}, Y = y^{(i)}) \\
= \ & P(X = x^{(i)} \mid Y = y^{(i)}) P(Y = y^{(i)}) \\
= \ & p(x^{(i)} \mid y^{(i)}) p(y^{(i)}) \\
= \ & p(x^{(i)} \mid y^{(i)}) p(y^{(i)})
\end{aligned}
$$

  - The probability of $\mathcal{D}$

$$
P(\mathcal{D}) = \prod_{i=1}^{m} P(X = x^{(i)}, Y = y^{(i)}) = \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) = \prod_{i=1}^{m} p(x^{(i)} \mid y^{(i)}) p(y^{(i)})
$$

Warm Up (Contd')

- Log-likelihood function

$$
\begin{aligned}
\ell(\theta) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) \\
&= \log \prod_{i=1}^{m} p(x^{(i)} \mid y^{(i)}) p(y^{(i)}) \\
&= \sum_{i=1}^{m} \left(\log p(x^{(i)} \mid y^{(i)}) + \log p(y^{(i)}) \right)
\end{aligned}
$$

where

$$
\theta = \{p_{(}x \mid y), p(y)\}_{x,y}
$$

Warm Up (Contd')

- Suppose we have n features

$$X = [X_1, X_2, \cdots, X_n]^T$$

- The features are independent with each other

$$
\begin{aligned}
P(X = x \mid Y = y) &= P(X_1 = x_1, \cdots, X_n = x_n \mid Y = y) \\
&= \prod_{j=1}^{n} P(X_j = x_j \mid Y = y) \\
&= \prod_{j=1}^{n} p_j(x_j \mid y)
\end{aligned}
$$

Warm Up (Contd')

- For each training data $(x^{(i)}, y^{(i)})$

$$
\begin{aligned}
& P(X = x^{(i)} \mid Y = y^{(i)}) \\
= \ & P(X_1 = x_1^{(i)}, \cdots, X_n = x_n^{(i)} \mid Y = y^{(i)}) \\
= \ & \prod_{j=1}^{n} P(X_j = x_j^{(i)} \mid Y = y^{(i)}) \\
= \ & \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y^{(i)})
\end{aligned}
$$

Gaussian Distribution

- Gaussian Distribution (Normal Distribution)

$$p(x; \mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

  where μ is the mean and σ^2 is the variance

- Gaussian distributions are important in statistics and are often used in the natural and social science to represent real-valued random variables whose distribution are not known

- Central limit theorem: The averages of samples of observations of random variables independently drawn from independent distributions converge in distribution to the normal, that is, become normally distributed when the number of observations is sufficiently large

  - Physical quantities that are expected to be the sum of many independent processes (such as measurement errors) often have distributions that are nearly normal.

Multivariate Gaussian Distribution

- Multivariate normal distribution in n-dimensions $\mathcal{N}(\mu, \Sigma)$

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

  - Mean vector $\mu \in \mathbb{R}^n$
  - Covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$
  - Mahalanobis distance: $r^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)$

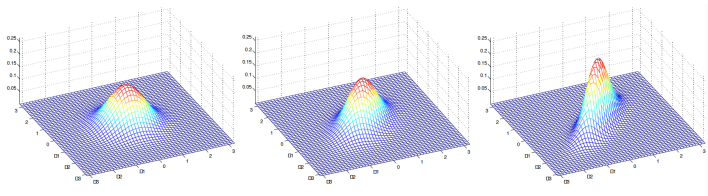- Σ is symmetric and positive semidefinite

$$\Sigma = \Phi \Lambda \Phi^T$$

  - Φ is an orthonormal matrix, whose columns are eigenvectors of Σ
  - Λ is a diagonal matrix with the diagonal elements being the eigenvalues
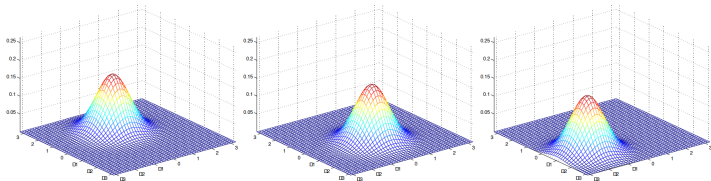
Multivariate Gaussian Distribution: A 2D Example
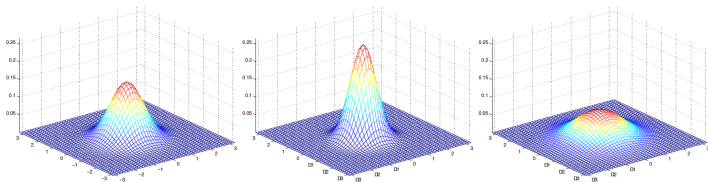
From left to right: $\Sigma = I$, $\Sigma = 0.6I$, $\Sigma = 2I$

From left to right: $\Sigma = I$, $\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$

Multivariate Gaussian Distribution: A 2D Example

From left to right: $\mu = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\mu = \begin{bmatrix} -0.5 \\ 0 \end{bmatrix}$, $\mu = \begin{bmatrix} -1 \\ -1.5 \end{bmatrix}$

Gaussian Discriminant Analysis (Contd.)

- $Y \sim Bernoulli(\psi)$
  - $P(Y = 1) = \psi$
  - $P(Y = 0) = 1 - \psi$
  - Probability mass function

$$p(y) = \psi^y (1 - \psi)^{1-y}, \quad \forall y = 0, 1$$

Gaussian Discriminant Analysis (Contd.)

- $X \mid Y = 0 \sim \mathcal{N}(\mu_0, \Sigma)$
  - Conditional probability density function of X given $Y = 0$

$$p_{X|Y=0}(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\right)$$

  - Or

$$p(x \mid y = 0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\right)$$

Gaussian Discriminant Analysis (Contd.)

- $X \mid Y = 1 \sim \mathcal{N}(\mu_1, \Sigma)$
  - Conditional probability density function of X given $Y = 1$

  $$p_{X\mid Y=1}(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right)$$

  - Or

  $$p(x \mid y = 1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right)$$

Gaussian Discriminant Analysis (Contd.)

- In summary, for $\forall y = 0, 1$

$$p(y) = \psi^y (1 - \psi)^{1-y}$$

$$p(x \mid y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_y)^T \Sigma^{-1}(x - \mu_y)\right)$$

Gaussian Discriminant Analysis (Contd.)

- Given m sample data, the log-likelihood is

$$
\begin{aligned}
&\ell(\psi, \mu_0, \mu_1, \Sigma) \\
=\ & \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \psi, \mu_0, \mu_1, \Sigma) \\
=\ & \log \prod_{i=1}^{m} p(x^{(i)} \mid y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \psi) \\
=\ & \sum_{i=1}^{m} \log p(x^{(i)} \mid y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^{m} \log p(y^{(i)}; \psi)
\end{aligned}
$$

Gaussian Discriminant Analysis (Contd.)

- The log-likelihood function

$$\ell(\psi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)} \mid y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^{m} \log p(y^{(i)}; \psi)$$

- For each training data $(x^{(i)}, y^{(i)})$
  - If $y^{(i)} = 0$

  $$p(x^{(i)} \mid y^{(i)}; \mu_0, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\right)$$

  - If $y^{(i)} = 1$

  $$p(x^{(i)} \mid y^{(i)}; \mu_1, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right)$$

Gaussian Discriminant Analysis (Contd.)

- The log-likelihood function

$$\ell(\psi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^{m} \log p(x^{(i)} \mid y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^{m} \log p(y^{(i)}; \psi)$$

- For each training data $(x^{(i)}, y^{(i)})$

$$p(y^{(i)}; \psi) = \psi^{y^{(i)}} (1 - \psi)^{1 - y^{(i)}}$$

Gaussian Discriminant Analysis (Contd.)

- Maximizing $\ell(\psi, \mu_0, \mu_1, \Sigma)$ through

$$\frac{\partial}{\partial \psi} \ell(\psi, \mu_0, \mu_1, \Sigma) = 0$$

$$\nabla_{\mu_0} \ell(\psi, \mu_0, \mu_1, \Sigma) = 0$$

$$\nabla_{\mu_1} \ell(\psi, \mu_0, \mu_1, \Sigma) = 0$$

$$\nabla_{\Sigma} \ell(\psi, \mu_0, \mu_1, \Sigma) = 0$$

Gaussian Discriminant Analysis (Contd.)

- Solutions:

$$\psi = \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\{y^{(i)} = 1\}$$

$$\mu_0 = \sum_{i=1}^{m}\mathbf{1}\{y^{(i)} = 0\}x^{(i)} / \sum_{i=1}^{m}\mathbf{1}\{y^{(i)} = 0\}$$

$$\mu_1 = \sum_{i=1}^{m}\mathbf{1}\{y^{(i)} = 1\}x^{(i)} / \sum_{i=1}^{m}\mathbf{1}\{y^{(i)} = 1\}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

- Proof (see Problem Set 2)

Gaussian Discriminant Analysis (Contd.)

- Given a test data sample x, we can calculate

$$
\begin{aligned}
p(y = 1 \mid x) &= \frac{p(x \mid y = 1)p(y = 1)}{p(x)} \\
&= \frac{p(x \mid y = 1)p(y = 1)}{p(x \mid y = 1)p(y = 1) + p(x \mid y = 0)p(y = 0)} \\
&= \frac{1}{1 + \frac{p(x|y=0)p(y=0)}{p(x|y=1)p(y=1)}}
\end{aligned}
$$

$$\frac{p(x \mid y = 0)p(y = 0)}{p(x \mid y = 1)p(y = 1)}$$

$$= \exp\left(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0) + \frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right) \cdot \frac{1 - \psi}{\psi}$$

$$= \exp\left((\mu_0 - \mu_1)^T \Sigma^{-1} x + \frac{1}{2}\left(\mu_1^T \Sigma^{-1}\mu_1 - \mu_0^T \Sigma^{-1}\mu_0\right)\right) \cdot \exp\left(\log\left(\frac{1 - \psi}{\psi}\right)\right)$$

$$= \exp\left((\mu_0 - \mu_1)^T \Sigma^{-1} x + \frac{1}{2}\left(\mu_1^T \Sigma^{-1}\mu_1 - \mu_0^T \Sigma^{-1}\mu_0\right)\right) \cdot \exp\left(\log\left(\frac{1 - \psi}{\psi}\right)\right)$$

$$= \exp\left((\mu_0 - \mu_1)^T \Sigma^{-1} x + \frac{1}{2}\left(\mu_1^T \Sigma^{-1}\mu_1 - \mu_0^T \Sigma^{-1}\mu_0\right) + \log\left(\frac{1 - \psi}{\psi}\right)\right)$$

- Assume

$$x := \begin{bmatrix} x \\ 1 \end{bmatrix}$$

$$\theta = \begin{bmatrix} (\mu_0 - \mu_1)^T \Sigma^{-1} \\ \frac{1}{2} \left(\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0 \right) + \log \left(\frac{1-\psi}{\psi} \right) \end{bmatrix}$$

- We have

$$\frac{p(x \mid y = 0)p(y = 0)}{p(x \mid y = 1)p(y = 1)}$$

$$= \exp \left((\mu_0 - \mu_1)^T \Sigma^{-1} x + \frac{1}{2} \left(\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0 \right) + \log \left(\frac{1-\psi}{\psi} \right) \right)$$

$$= \exp \left(\theta^T x \right)$$

- Then

$$p(y = 1 \mid x) = \frac{1}{1 + \frac{p(x|y=0)p(y=0)}{p(x|y=1)p(y=1)}} = \frac{1}{1 + \exp(\theta^T x)}$$

- Similarly, we have

$$p(y = 0 \mid x)$$

$$= \frac{p(x \mid y = 0)p(y = 0)}{p(x)}$$

$$= \frac{p(x \mid y = 0)p(y = 0)}{p(x \mid y = 1)p(y = 1) + p(x \mid y = 0)p(y = 0)}$$

$$= \frac{1}{1 + \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)}}$$

$$= \frac{1}{1 + \exp\left((\mu_1 - \mu_0)^T \Sigma^{-1} x + \frac{\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1}{2} + \log\left(\frac{\psi}{1 - \psi}\right)\right)}$$

GDA and Logistic Regression

- GDA can model be reformulated as logistic regression
- Which one is better?
  - GDA makes stronger modeling assumptions, and is more data efficient (i.e., requires less training data to learn "well") when the modeling assumptions are correct or at least approximately correct
  - Logistic regression makes weaker assumptions, and is significantly more robust deviations from modeling assumptions
  - When the data is indeed non-Gaussian, then in the limit of large datasets, logistic regression will almost always do better than GDA
  - In practice, logistic regression is used more often than GDA

Gaussian Discriminant Analysis (Contd.)
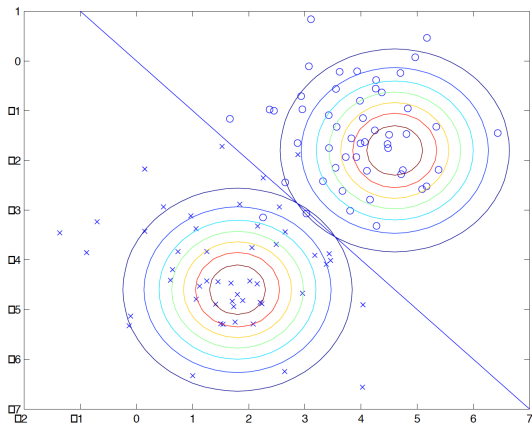
Naive Bayes

- Training data $(x^{(i)}, y^{(i)})_{i=1,\cdots,m}$
    - $x^{(i)}$ is a n-dimensional vector
    - Each feature $x_j^{(i)} \in \{0, 1\}$ $(j = 1, \cdots, n)$
    - $y^{(i)} \in \{1, \cdots, k\}$
- The features and labels can be represented by random variables $\{X_j\}_{j=1,\cdots,n}$ and Y, respectively

Naive Bayes (Contd')

- For $\forall j \neq j'$, Naive Bayes assumes X_j and $X_{j'}$ are conditionally independent given Y

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n \mid Y = y)$$

$$= \prod_{j=1}^{n} P(X_j = x_j \mid X_1 = x_1, X_2 = x_2, \cdots, X_{j-1} = x_{j-1}, Y = y)$$

$$= \prod_{j=1}^{n} P(X_j = x_j \mid Y = y)$$

Naive Bayes (Contd.)

- The key assumption in NB model

$$
\begin{aligned}
& P(Y = y, X_1 = x_1, \cdots, X_n = x_n) \\
= \; & P(X_1 = x_1, \cdots, X_n = x_n \mid Y = y)P(Y = y) \\
= \; & P(Y = y) \prod_{j=1}^{n} P(X_j = x_j \mid Y = y)
\end{aligned}
$$

- The key assumption in NB model

$$
\begin{aligned}
P(Y = y, X_1 = x_1, \cdots, X_n = x_n) &= P(Y = y) \prod_{j=1}^{n} P(X_j = x_j \mid Y = y) \\
&= p(y) \prod_{j=1}^{n} p_j(x_j \mid y)
\end{aligned}
$$

- Two sets of parameters (denoted by Ω)

  - Probability mass function of Y

  $$
  p(y) = P(Y = y)
  $$

  where $\forall y \in \{1, 2, \cdots, k\}$
  - Conditional probability mass function of X_j ($j \in \{1, 2, \cdots, n\}$) given $Y = y$ ($y \in \{1, \cdots, k\}$)
  $$
  p_j(x \mid y) = P(X_j = x \mid Y = y)
  $$

  where $\forall x_j \in \{0, 1\}$

Maximum-Likelihood Estimates for Naive Bayes

- Log-likelihood function is

$$
\begin{aligned}
\ell(\Omega) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) \\
&= \sum_{i=1}^{m} \log p(x^{(i)}, y^{(i)}) \\
&= \sum_{i=1}^{m} \log \left(p(y^{(i)}) \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y^{(i)}) \right) \\
&= \sum_{i=1}^{m} \log p(y^{(i)}) + \sum_{i=1}^{m} \sum_{j=1}^{n} \log p_j(x_j^{(i)} \mid y^{(i)})
\end{aligned}
$$

MLE for Naive Bayes (Contd.)

- Notation:
    - The number of training data whose label is y

    $$count(y) = \sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y), \ \forall y = 1, 2, \cdots, k$$

    - The number of training data with the j-th feature being x and the label being y

    $$count_j(x \mid y) = \sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y \wedge x_j^{(i)} = x), \ \forall y = 1, 2, \cdots, k, \ \forall x = 0, 1$$

MLE for Naive Bayes (Contd.)

- The solutions to maximizing the log-likelihood function

$$p(y) = \frac{count(y)}{m} = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y)}{m}$$

and

$$p_j(x \mid y) = \frac{count_j(x \mid y)}{count(y)} = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y \wedge x_j^{(i)} = x)}{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y)}$$

$$\sum_{i=1}^{m} \log p(y^{(i)})$$
$$= \sum_{i:y^{(i)}=1} \log p(y=1) + \cdots + \sum_{i:y^{(i)}=k} \log p(y=k)$$
$$= count(y=1) \log p(y=1) + \cdots + count(y=k) \log p(y=k)$$
$$= \sum_{y=1}^{k} count(y) \log p(y)$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \log p_j(x_j^{(i)} \mid y^{(i)})$$

$$= \sum_{j=1}^{n} \sum_{i:y^{(i)}=1} \log p_j(x_j^{(i)} \mid y = 1) + \cdots + \sum_{j=1}^{n} \sum_{i:y^{(i)}=k} \log p_j(x_j^{(i)} \mid y = k)$$

- $D_{x|y}^{(j)} = \{i : x_j^{(i)} = x \land y^{(i)} = y\}$
- Then, we have

$$
\sum_{i=1}^{m} \sum_{j=1}^{n} \log p_j(x_j^{(i)} \mid y^{(i)})
$$

$$
= \sum_{j=1}^{n} \sum_{i:y^{(i)}=1} \log p_j(x_j^{(i)} \mid y = 1) + \cdots + \sum_{j=1}^{n} \sum_{i:y=k} \log p_j(x_j^{(i)} \mid y = k)
$$

$$
= \sum_{j=1}^{n} \left(\sum_{i \in \mathcal{D}_{0|1}^{(j)}} \log p_j(x = 0 \mid y = 1) + \sum_{i \in \mathcal{D}_{1|1}^{(j)}} \log p_j(x = 1 \mid y = 1) \right)
$$

$$
\vdots
$$

$$
+ \sum_{j=1}^{n} \left(\sum_{i \in \mathcal{D}_{0|k}^{(j)}} \log p_j(x = 0 \mid y = k) + \sum_{i \in \mathcal{D}_{1|k}^{(j)}} \log p_j(x = 1 \mid y = k) \right)
$$

- $D_{x|y}^{(j)} = \{i : x_j^{(i)} = x \land y^{(i)} = y\}$
- Then, we have

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \log p_j(x_j^{(i)} \mid y^{(i)})$$

$$= \sum_{j=1}^{n} \left(\sum_{i \in \mathcal{D}_{0|1}^{(j)}} \log p_j(x = 0 \mid y = 1) + \sum_{i \in \mathcal{D}_{1|1}^{(j)}} \log p_j(x = 1 \mid y = 1) \right)$$

$$\vdots$$

$$+ \sum_{j=1}^{n} \left(\sum_{i \in \mathcal{D}_{0|k}^{(j)}} \log p_j(x = 0 \mid y = k) + \sum_{i \in \mathcal{D}_{1|k}^{(j)}} \log p_j(x = 1 \mid y = k) \right)$$

$$= \sum_{j=1}^{n} \sum_{y=1}^{k} \sum_{x \in \{0,1\}} count_j(x \mid y) \log p_j(x \mid y)$$

MLE for Naive Bayes (Contd.)

- Revisiting the log-likelihood function

$$
\begin{aligned}
\ell(\Omega) &= \sum_{i=1}^{m} \log p(y^{(i)}) + \sum_{i=1}^{m} \sum_{j=1}^{n} \log p_j(x_j^{(i)} \mid y^{(i)}) \\
&= \sum_{y=1}^{k} count(y) \log p(y) + \sum_{j=1}^{n} \sum_{y=1}^{k} \sum_{x \in \{0,1\}} count_j(x \mid y) \log p_j(x \mid y)
\end{aligned}
$$

MLE for Naive Bayes (Contd.)

- Assume a training set $\{(x^{(i)}, y^{(i)})\}_{i=1,\cdots,m}$. The maximum-likelihood estimates are then the parameter values $p(y)$ and $p_j(x_j|y)$ that maximize

$$\ell(\Omega) = \sum_{y=1}^{k} count(y) \log p(y) + \sum_{j=1}^{n} \sum_{y=1}^{k} \sum_{x \in \{0,1\}} count_j(x \mid y) \log p_j(x \mid y)$$

subject to the following constraints:
- $p(y) \geq 0$ for $\forall y \in \{1, 2, \cdots, k\}$
- $\sum_{y=1}^{k} p(y) = 1$
- For $\forall y \in \{1, \cdots, k\}, j \in \{1, \cdots, n\}, x \in \{0, 1\}$, $p_j(x \mid y) \geq 0$
- For $\forall y \in \{1, \cdots, k\}, j \in \{1, \cdots, n\}$, $\sum_{x \in \{0,1\}} p_j(x \mid y) = 1$

MLE for Naive Bayes (Contd.)

- **Theorem 1**
  The maximum-likelihood estimates for Naive Bayes model are as follows

  $$p(y) = \frac{count(y)}{m} = \frac{\sum_{i=1}^m \mathbf{1}(y^{(i)} = y)}{m}$$

  and

  $$p_j(x \mid y) = \frac{count_j(x \mid y)}{count(y)} = \frac{\sum_{i=1}^m \mathbf{1}(y^{(i)} = y \wedge x_j^{(i)} = x)}{\sum_{i=1}^m \mathbf{1}(y^{(i)} = y)}$$

- Proof (see Problem Set 2)

- What if $x \in \{1, 2, \cdots, u\}$ instead of $x \in \{0, 1\}$?
- Can we get the same results? Check it yourself!

Laplace Smoothing

- There may exist some feature, e.g., X_j, such that $X_j = x$ may never happen in the training data

$$p_j(x|y) = 0 \quad \Rightarrow \quad p(y|x_1, \cdots, x_n) = 0$$

- This is unreasonable!!!
- How can we resolve this problem?
    - Laplace smoothing:

$$p(y) = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y) + 1}{m + k}$$

$$p_j(x \mid y) = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y \wedge x_j^{(i)} = x) + 1}{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y) + v_j}$$

where v is the number of the possible values of the j-th feature ($v_j = 2$ for $\forall j = 1, \cdots, n$ in our case)

Classification by Naive Bayes

- Given a test sample $x = [x_1, x_2, \cdots, x_3]^T$, we have

$$
\begin{aligned}
& P(Y = y \mid X_1 = x_1, \cdots, X_n = x_n) \\
= {} & \frac{P(X_1 = x_1, \cdots, X_n = x_n \mid Y = y)P(Y = y)}{P(X_1 = x_1, \cdots, X_n = x_n)} \\
= {} & \frac{P(Y = y)\prod_{j=1}^n p(X_j = x_j \mid Y = y)}{P(X_1 = x_1, \cdots, X_n = x_n)} \\
= {} & \frac{p(y)\prod_{j=1}^n p_j(x_j \mid y)}{p(x_1, \cdots, x_n)}
\end{aligned}
$$

- Therefore, the output of the Naive Bayes model is

$$
\arg \max_{y \in \{1, \cdots, k\}} \left(p(y) \prod_{j=1}^n p_j(x_j \mid y) \right)
$$

Classification by Naive Bayes

- Example: $y = 0, 1, 2, 3$

$$P(Y = 0 \mid X_1 = x_1, \cdots, X_n = x_n) = \frac{p(y=0) \prod_{j=1}^{n} p_j(x_j \mid y = 0)}{p(x_1, \cdots, x_n)}$$

$$P(Y = 1 \mid X_1 = x_1, \cdots, X_n = x_n) = \frac{p(y=1) \prod_{j=1}^{n} p_j(x_j \mid y = 1)}{p(x_1, \cdots, x_n)}$$

$$P(Y = 2 \mid X_1 = x_1, \cdots, X_n = x_n) = \frac{p(y=2) \prod_{j=1}^{n} p_j(x_j \mid y = 2)}{p(x_1, \cdots, x_n)}$$

$$P(Y = 3 \mid X_1 = x_1, \cdots, X_n = x_n) = \frac{p(y=3) \prod_{j=1}^{n} p_j(x_j \mid y = 3)}{p(x_1, \cdots, x_n)}$$

Naive Bayes for Multinomial Distribution

- Assumptions:
  - Each training sample involves a different number of features

$$x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \cdots, x_{n_i}^{(i)}]^{\mathrm{T}}$$

  - The j-th feature of $x^{(i)}$ takes a finite set of values

$$x_j^{(i)} \in \{1, 2, \cdots, v\}, \text{ for } \forall j = 1, \cdots, n$$

  - For each training data, the features are i.i.d.

$$P(X_j = t \mid Y = y) = p(t \mid y), \text{ for } \forall j = 1, \cdots, n$$

    - $p(t \mid y) \geq 0$ is the conditional probability mass function
    - $\sum_{t=1}^{v} p(t \mid y) = 1$

Naive Bayes for Multinomial Distribution (Contd.)

- Assumptions:
  - Each training sample involves a different number of features

  $$x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \cdots, x_{n_i}^{(i)}]^{\mathrm{T}}$$

  - The j-th feature of $x^{(i)}$ takes a finite set of values, $x_j^{(i)} \in \{1, 2, \cdots, v\}$
- For each training data (x, y) where x is a n-dimensional vector

  $$P(Y = y) = p(y)$$

  $$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n \mid Y = y) = \prod_{j=1}^{n} P(X_j = x_j \mid Y = y)$$

Naive Bayes for Multinomial Distribution (Contd.)

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n \mid Y = y)$$

$$= \prod_{j=1}^{n} P(X_j = x_j \mid Y = y)$$

$$= \prod_{t=1}^{v} \prod_{j:X_j=t} P(X_j = t \mid Y = y)$$

$$= \prod_{t=1}^{v} \prod_{j:X_j=t} p(t \mid y)$$

$$= \prod_{t=1}^{v} p(t \mid y)^{count(t)}$$

where

$$count(t) = \sum_{j=1}^{n} \mathbf{1}(x_j = t)$$

Naive Bayes for Multinomial Distribution (Contd.)

- For each training data (x, y)

$$P(Y = y) = p(y)$$

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n \mid Y = y) = \prod_{t=1}^{v} p(t \mid y)^{count(t)}$$

- Parameters Ω

$$p(y), \ y = 1, 2, \cdots, k$$
$$p(t \mid y), \ t = 1, 2, \cdots, v, \ y = 1, 2, \cdots, k$$

Naive Bayes for Multinomial Distribution (Contd.)

- Log-likelihood function

$$
\begin{aligned}
\ell(\Omega) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) \\
&= \log \prod_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) p(x^{(i)} \mid y^{(i)} = y) p(y^{(i)} = y) \\
&= \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) \log \left(p(x^{(i)} \mid y^{(i)} = y) p(y^{(i)} = y) \right)
\end{aligned}
$$

Naive Bayes for Multinomial Distribution (Contd.)

$$
\begin{aligned}
\ell(\Omega) &= \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) \log \Big(p(x^{(i)} \mid y^{(i)} = y) p(y^{(i)} = y) \Big) \\
&= \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) \log \left(p(y^{(i)} = y) \prod_{j=1}^{n_i} p(x_j^{(i)} \mid y^{(i)} = y) \right) \\
&= \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) \log \left(p(y) \prod_{t=1}^{v} p(t \mid y)^{count^{(i)}(t)} \right) \\
&= \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) \left(\log p(y) + \sum_{t=1}^{v} count^{(i)}(t) p(t \mid y) \right)
\end{aligned}
$$

where

$$
count^{(i)}(t) = \sum_{j=1}^{n_i} \mathbf{1}(x_j^{(i)} = t)
$$

Naive Bayes for Multinomial Distribution (Contd.)

- Problem formulation

$$\max \quad \ell(\Omega) = \sum_{i=1}^{m} \sum_{y=1}^{k} \mathbf{1}(y^{(i)} = y) \left(\log p(y) + \sum_{t=1}^{v} count^{(i)}(t) p(t \mid y) \right)$$

$$s.t. \quad p(y) \geq 0, \ \forall y = 1, \cdots, k$$

$$p(t \mid y) \geq 0, \ \forall t = 1, \cdots, v \ \forall y = 1, \cdots, k$$

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

$$\sum_{t=1}^{v} p(t \mid y) = 1, \ \forall y = 1, 2, \cdots, k$$

Naive Bayes for Multinomial Distribution (Contd.)

- Solution

$$p(t \mid y) = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y) count^{(i)}(t)}{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y) \sum_{t=1}^{v} count^{(i)}(t)}$$

$$p(y) = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y)}{m}$$

- Check them by yourselves!

Naive Bayes for Multinomial Distribution (Contd.)

- Laplace smoothing

$$\psi(t \mid y) = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y) count^{(i)}(t)}{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y) \sum_{t=1}^{v} count^{(i)}(t) + v}$$

$$\psi(y) = \frac{\sum_{i=1}^{m} \mathbf{1}(y^{(i)} = y)}{m + k}$$

Convex Functions

- A set C is convex if the line segment between any two points in C lies in C, i.e., for $\forall x_1, x_2 \in C$ and $\forall \theta$ with $0 \leq \theta \leq 1$, we have

$$\theta x_1 + (1 - \theta) x_2 \in C$$

- A function $f : \mathbb{R}^n \to R$ is convex, if $\mathbf{dom} f$ is a convex set and if for all $x, y \in \mathbf{dom} f$ and λ with $0 \leq \lambda \leq 1$, we have

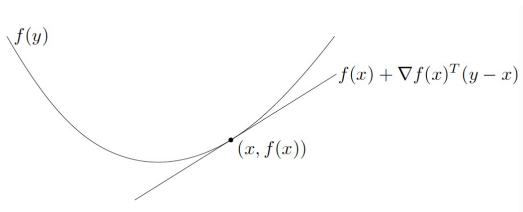$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y)$$

Convex Functions (Contd.)

- First-order conditions: Suppose f is differentiable (i.e., its gradient ∇f exists at each point in $\mathbf{dom}f$, which is open). Then, f is convex if and only if $\mathbf{dom}f$ is convex and

$$f(y) \geq f(x) + \nabla f(x)^T(y - x)$$

holds for $\forall x, y \in \mathbf{dom}f$

Convex Functions (Contd.)

- Second-order conditions: Assume f is twice differentiable (i.t., its Hessian matrix or second derivative $\nabla^2 f$ exists at each point in $\mathbf{dom} f$, which is open), then f is convex if and only if $\mathbf{dom} f$ is convex and its Hessian is positive semidefinite: for $\forall x \in \mathbf{dom} f$,

$$\nabla^2 f \succeq 0$$

Jensen's Inequality

- Let f be a convex function, then

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2)$$

where $\lambda \in [0, 1]$

- Jensen's Inequality
  - Let $f(x)$ be a convex function defined on an interval $\mathcal{I}$. If $x_1, x_2, \cdots, x_N \in \mathcal{I}$ and $\lambda_1, \lambda_2, \cdots, \lambda_N \geq 0$ with $\sum_{i=1}^{N} \lambda_i = 1$

$$f(\sum_{i=1}^{N} \lambda_i x_i) \leq \sum_{i=1}^{N} \lambda_i f(x_i)$$

The Proof of Jensen's Inequality

- When $N = 1$, the result is trivial
- When $N = 2$,

$$f(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2)$$

  due to convexity of $f(x)$
- When $N \geq 3$, the proof is by induction
  - We assume that, the Jensen's inequality holds when $N = k - 1$.

  $$f(\sum_{i=1}^{k-1} \lambda_i x_i) \leq \sum_{i=1}^{k-1} \lambda_i f(x_i)$$

  - We then prove that, the Jensen's inequality still holds for $N = k$

- When $N = k$,

$$
\begin{aligned}
f(\sum_{i=1}^{k} \lambda_i x_i) &= f(\sum_{i=1}^{k-1} \lambda_i x_i + \lambda_k x_k) \\
&= f((1 - \lambda_k) \sum_{i=1}^{k-1} \frac{\lambda_i}{1 - \lambda_k} x_i + \lambda_k x_k) \\
&\leq (1 - \lambda_k) f(\sum_{i=1}^{k-1} \frac{\lambda_i}{1 - \lambda_k} x_i) + \lambda_k f(x_k) \\
&\leq (1 - \lambda_k) \sum_{i=1}^{k-1} \frac{\lambda_i}{1 - \lambda_k} f(x_i) + \lambda_k f(x_k) \\
&= \sum_{i=1}^{k-1} \lambda_i f(x_i) + \lambda_k f(x_k) \\
&= \sum_{i=1}^{k} \lambda_i f(x_i)
\end{aligned}
$$

The Probabilistic Form of Jensen's Inequality

- The inequality can be extended to infinite sums, integrals, and expected values
- If $p(x) \geq 0$ on $\mathcal{S} \subseteq \mathbf{dom} f$ and $\int_{\mathcal{S}} p(x) dx = 1$, we have

$$f(\int_{\mathcal{S}} p(x) x dx) = \int_{\mathcal{S}} p(x) f(x) dx$$

- Assuming X is a random variable and P is a probability distribution on sample space $\mathcal{S}$, we have

$$f[E(X)] \leq E[f(X)]$$

- The equality holds if X is a constant

Jensen's inequality for Concave Function

- Assume f be a concave function

$$f(\sum_{i=1}^{N} \lambda_i x_i) \geq \sum_{i=1}^{N} \lambda_i f(x_i)$$

- The probabilistic form

$$f(E[X]) \geq E[f(X)]$$

- Example: $f(x) = \log x$
  - $\log(\sum_{i=1}^{N} \lambda_i x_i) \geq \sum_{i=1}^{N} \lambda_i \log(x_i)$

  - $\log(E[X]) \geq E[\log(X)]$

The Expectation-Maximization (EM) Algorithm

- A training set $\{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$ (without labels)
- The log-likelihood function

$$
\begin{aligned}
\ell(\theta) &= \log \prod_{i=1}^{m} p(x^{(i)}; \theta) \\
&= \sum_{i=1}^{m} \log \sum_{z^{(i)} \in \Omega} p(x^{(i)}, z^{(i)}; \theta)
\end{aligned}
$$

- θ denotes the full set of unknown parameters in the model
- $z^{(i)} \in \Omega$ is so-called "latent variable"

The EM Algorithm (Contd')

- Our goal is to maximize the log-likelihood function

$$\ell(\theta) = \sum_{i=1}^{m} \log \sum_{z^{(i)} \in \Omega} p(x^{(i)}, z^{(i)}; \theta)$$

- The basic idea of EM algorithm
  - Repeatedly construct a lower-bound on ℓ (E-step)
  - Then optimize that lower-bound (M-step)

The EM Algorithm (Contd')

- A training set $\{x^{(1)}, x^{(2)}, \cdots, x^{(m)}\}$ (without labels)
- The log-likelihood function

$$\ell(\theta) = \sum_{i=1}^{m} \log p(x^{(i)}; \theta) = \sum_{i=1}^{m} \log \sum_{z^{(i)} \in \Omega} p(x^{(i)}, z^{(i)}; \theta)$$

  - θ denotes the full set of unknown parameters in the model
  - $z^{(i)} \in \Omega$'s are latent variables
- Then, our goal is to maximize the above log-likelihood function
- The basic idea of EM algorithm: Repeatedly construct a lower-bound on ℓ (E-step), and then optimize that lower-bound (M-step)

- For each $i \in \{1, 2, \cdots, m\}$, let $Z_i \in \Omega$ be a random variable indicating the label of the i-th training data

- The probability that the label of the i-th training data is z:

$$P(Z_i = z) = Q_i(z)$$

- We say Q_i be some distribution over the z's

$$\sum_{z \in \Omega} Q_i(z) = 1, \quad Q_i(z) \geq 0$$

- We have

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{m} \log \sum_{z^{(i)} \in \Omega} p(x^{(i)}, z^{(i)}; \theta) \\
&= \sum_{i=1}^{m} \log \sum_{z^{(i)} \in \Omega} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}
\end{aligned}
$$

- Assume $\phi : \Omega \to R$

$$
\begin{aligned}
E[\phi(Z_i)] &= \sum_{z^{(i)} \in \Omega} P(Z_i = z^{(i)}) \phi(z^{(i)}) \\
&= \sum_{z^{(i)} \in \Omega} Q_i(z^{(i)}) \phi(z^{(i)})
\end{aligned}
$$

- Suppose

$$
\phi(z^{(i)}) = \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}
$$

- Re-visit the log-likelihood function

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{m} \log \sum_{z^{(i)} \in \Omega} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \\
&= \sum_{i=1}^{m} \log E[\phi(Z_i)]
\end{aligned}
$$

- Since $\log(\cdot)$ is a concave function, according to Jensen's inequality, we have

$$\log(E[\phi(Z_i)]) \geq E[\log(\phi(Z_i))]$$

- Then, the log-likelihood function

$$
\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{m} \log E[\phi(Z_i)] \\
&\geq \sum_{i=1}^{m} E[\log(\phi(Z_i)] \\
&= \sum_{i=1}^{m} \sum_{z^{(i)} \in \Omega} Q_i(z^{(i)}) \log \phi(z^{(i)}) \\
&= \sum_{i=1}^{m} \sum_{z^{(i)} \in \Omega} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}
\end{aligned}
$$

- For any set of distributions Q_i, $\ell(\theta)$ has a lower bound

$$\ell(\theta) \geq \sum_{i=1}^{m} \sum_{z^{(i)} \in \Omega} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

- Tighten the lower bound (i.e., let the equality hold)
  - The equality in the Jensen's inequality holds if

$$\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} = c$$

  where c is a constant

- Tighten the lower bound (i.e., let the equality hold)
  - The equality in the Jensen's inequality holds if

$$\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} = c$$

  where c is a constant
  - Since

$$\sum_{z^{(i)} \in \Omega} Q_i(z^{(i)}) = 1$$

  we have

$$\sum_{z^{(i)} \in \Omega} p(x^{(i)}, z^{(i)}; \theta) = c \sum_{z^{(i)} \in \Omega} Q_i(z) = c$$

- Tighten the lower bound (i.e., let the equality hold)
    - We have

$$\begin{cases} p(x^{(i)}, z^{(i)}; \theta)/Q_i(z^{(i)}) = c \\ \sum_{z^{(i)} \in \Omega} Q_i(z) = 1 \\ \sum_{z^{(i)} \in \Omega} p(x^{(i)}, z^{(i)}; \theta) = c \end{cases}$$

    - Therefore,

$$\begin{aligned} Q_i(z^{(i)}) &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{c} \\ &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{\sum_{z^{(i)} \in \Omega} p(x^{(i)}, z^{(i)}; \theta)} \\ &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} \\ &= p(z^{(i)} \mid x^{(i)}; \theta) \end{aligned}$$

The EM Algorithm (Contd')

- Repeat the following step until convergence
    - (E-step) For each i, set

    $$Q_i(z^{(i)}) := p(z^{(i)} \mid x^{(i)}; \theta)$$

    - (M-step) set

    $$\theta := \arg\max_{\theta} \sum_i \sum_{z^{(i)} \in \Omega} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})}$$

Convergence

- In the end of the t-th iteration,

$$\ell(\theta^{[t]}) = \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i^{[t]}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{[t]})}{Q_i^{[t]}(z^{(i)})}$$

- $\theta^{(t+1)}$ is then obtained by maximizing the right hand side of the above equation

$$
\begin{aligned}
\ell(\theta^{[t+1]}) &\geq \sum_{i=1}^{m} \sum_{z^{(i)} \in \Omega} Q_i^{[t]}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{[t+1]})}{Q_i^{[t]}(z^{(i)})} \\
&\geq \sum_{i=1}^{m} \sum_{z^{(i)} \in \Omega} Q_i^{[t]}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{[t]})}{Q_i^{[t]}(z^{(i)})} \\
&= \ell(\theta^{[t]})
\end{aligned}
$$

- EM causes the likelihood to converge monotonically

Reviewing Mixtures of Gaussians

- A training set $\{x^{(1)}, \cdots, x^{(m)}\}$
- Mixture of Gaussians model

$$p(x, z) = p(x \mid z)p(z)$$

  - $Z \in \{1, \cdots, k\} \sim \text{Multinomial}(\{\phi_j\}_{j=1,\cdots,k})$
  - $\phi_j = P(Z = j)$ ($\phi_j \geq 0$ and $\sum_{j=1}^{k} \phi_j = 1$)
  - $x \mid z = j \sim \mathcal{N}(\mu_j, \Sigma_j)$
  - z's are so-called latent random variables, since they are hidden/unobserved

Reviewing Mixtures of Gaussians (Contd.)

- The log-likelihood function

$$
\begin{aligned}
\ell(\phi, \mu, \Sigma) &= \sum_{i=1}^{m} \log p(x^{(i)}; \phi, \mu, \Sigma) \\
&= \sum_{i=1}^{m} \log \sum_{z^{(i)}=1}^{k} p(x^{(i)} \mid z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi)
\end{aligned}
$$

Applying EM Algorithm to Mixtures of Gaussians

- Repeat the following steps until convergence
    - (E-step) For each i, j, set

$$
\begin{aligned}
w_j^{(i)} &= p(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \Sigma) \\
&= \frac{p(x^{(i)} \mid z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)} \mid z^{(i)} = l; \mu, \Sigma) p(z^{(i)} = l; \phi)}
\end{aligned}
$$

    - (M-step) Update the parameters

$$
\phi_j = \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)}
$$

$$
\mu_j = \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}}
$$

$$
\Sigma_j = \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}
$$

Applying EM Algorithm to MG (Contd.)

- (E-step) For each i, j, set

$$
\begin{aligned}
w_j^{(i)} &= Q_i(z^{(i)} = j) \\
&= p(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \Sigma) \\
&= \frac{p(x^{(i)} \mid z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)} \mid z^{(i)} = l; \mu, \Sigma) p(z^{(i)} = l; \phi)}
\end{aligned}
$$

  - $p(x^{(i)} \mid z^{(i)} = j; \mu_j, \Sigma_j) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1}(x - \mu_j)\right)$
  - $p(z^{(i)} = j; \phi) = \phi_j$

Applying EM Algorithm to MG (Contd.)

- (M-step) Maximizing

$$\sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{k} Q_i(z^{(i)} = j) \log \frac{p(x^{(i)} \mid z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{Q_i(z^{(i)} = j)}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp \left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right) \phi_j}{\omega_j^{(i)}}$$

$$= -\sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} \left[\log \left((2\pi)^{\frac{n}{2}} |\Sigma_j|^{\frac{1}{2}} \right) + \frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right]$$

$$\quad + \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} \log \phi_j - \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} \log \omega_j^{(i)}$$

Applying EM Algorithm to MG: Calculating μ

$$\nabla_{\mu_l} \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(j)} - \mu_j)^T \Sigma_j^{-1}(x^{(j)} - \mu_j)\right) \phi_j}{\omega_j^{(i)}}$$

$$= \nabla_{\mu_l} \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} \frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1}(x^{(i)} - \mu_j)$$

$$= \frac{1}{2} \sum_{i=1}^{m} \omega_l^{(i)} \nabla_{\mu_l} \left(2\mu_l^T \Sigma_l^{-1} x^{(i)} - \mu_l^T \Sigma_l^{-1} \mu_l\right)$$

$$= \sum_{i=1}^{m} \omega_l^{(i)} \left(\Sigma_l^{-1} x^{(i)} - \Sigma_l^{-1} \mu_l\right)$$

Let $\sum_{i=1}^{m} \omega_l^{(i)} \left(\Sigma_l^{-1} x^{(i)} - \Sigma_l^{-1} \mu_l\right) = 0$, we have

$$\mu_l = \frac{\sum_{i=1}^{m} \omega_l^{(i)} x^{(i)}}{\sum_{i=1}^{m} \omega_l^{(i)}}$$

Applying EM Algorithm to MG: Calculating ϕ

- Our problem becomes

$$\max \ \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} \log \phi_j$$

$$\text{s.t.} \ \sum_{j=1}^{k} \phi_j = 1$$

- Using the theory of Lagrange multiplier

$$\phi_j = \frac{1}{m} \sum_{i=1}^{m} \omega_j^{(i)}$$

Applying EM Algorithm to MG: Calculating Σ

- Minimizing

$$\sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} \left[\log \left((2\pi)^{\frac{n}{2}} |\Sigma_j|^{\frac{1}{2}} \right) + \frac{1}{2} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right]$$

$$= \frac{n \log 2\pi}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} \log |\Sigma_j|$$

$$+ \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)$$

- Minimizing

$$\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} \log |\Sigma_j| + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)$$

- Therefore, we have

$$\nabla_{\Sigma_j} \left(\sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} \log |\Sigma_j| + \sum_{i=1}^{m} \sum_{j=1}^{k} \omega_j^{(i)} (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j) \right) = 0$$

for $\forall j \in \{1, \cdots, k\}$

- By applying

$$\nabla_X \text{tr}(AX^{-1}B) = -(X^{-1}BAX^{-1})^T$$
$$\nabla_A |A| = |A|(A^{-1})^T$$

- We get a solution

$$\Sigma_j = \frac{\sum_{i=1}^m \omega_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^m \omega_j^{(i)}}$$

where $j = 1, \cdots, k$

- Check the derivations by yourself!

Naive Bayes with Missing Labels

- For any x, we have

$$p(x) = \sum_{y=1}^{k} p(x,y) = \sum_{y=1}^{k} \left(p(y) \prod_{j=1}^{n} p_j(x_j \mid y) \right)$$

- The log-likelihood function is then defined as

$$\ell(\theta) = \sum_{i=1}^{m} \log p(x^{(i)}) = \sum_{i=1}^{m} \log \sum_{y=1}^{k} \left(p(y) \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y) \right)$$

- Maximizing $\ell(\theta)$ subject to the following constraints
  - $p(y) \geq 0$ for $\forall y \in \{1, \cdots, k\}$, and $\sum_{y=1}^{k} p(y) = 1$
  - For $\forall y \in \{1, \cdots, k\}, j \in \{1, \cdots, n\}, x_j \in \{0,1\}$, $p_j(x_j \mid y) \geq 0$
  - For $\forall y \in \{1, \cdots, k\}$ and $j \in \{1, \cdots, n\}$, $\sum_{x_j \in \{0,1\}} p_j(x_j \mid y) = 1$

Naive Bayes with Missing Labels (Contd.)

- When labels are given

$$\ell(\theta) = \sum_{i=1}^{m} \log \left(p(y^{(i)}) \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y^{(i)}) \right)$$

- When labels are missed

$$\ell(\theta) = \sum_{i=1}^{m} \log \sum_{y=1}^{k} \left(p(y) \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y) \right)$$

Applying EM Algorithm to Naive Bayes

- Repeat the following steps until convergence
  - (E-step) For each $i = 1, \cdots, m$ and $y = 1, \cdots, k$ set

  $$\delta(y \mid i) = p(y \mid x^{(i)}) = \frac{p(y) \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y)}{\sum_{y'=1}^{k} p(y') \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y')}$$

  - (M-step) Update the parameters

  $$p(y) = \frac{1}{m} \sum_{i=1}^{m} \delta(y \mid i)$$

  $$p_j(x \mid y) = \frac{\delta(y \mid i)}{\sum_{i=1}^{m} \delta(y \mid i)}$$

Applying EM Algorithm to NB: E-Step

$$
\begin{aligned}
\delta(y \mid i) &= p(y \mid x^{(i)}) \\
&= \frac{p(x^{(i)} \mid y)p(y)}{p(x^{(i)})} \\
&= \frac{p(y) \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y)}{\sum_{y'=1}^{k} p(x^{(i)}, y')} \\
&= \frac{p(y) \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y)}{\sum_{y'=1}^{k} p(x^{(i)} \mid y')p(y')} \\
&= \frac{p(y) \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y)}{\sum_{y'=1}^{k} p(y') \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y')}
\end{aligned}
$$

Applying EM Algorithm to NB: M-Step

$$\sum_{i=1}^{m} \sum_{y=1}^{k} \delta(y \mid i) \log \frac{p(x^{(i)}, z^{(i)} = y)}{\delta(y \mid i)}$$

$$= \sum_{i=1}^{m} \sum_{y=1}^{k} \delta(y \mid i) \log \frac{p(x^{(i)} \mid z^{(i)} = y) p(z^{(i)} = y)}{\delta(y \mid i)}$$

$$= \sum_{i=1}^{m} \sum_{y=1}^{k} \delta(y \mid i) \log \frac{p(y) \prod_{j=1}^{n} p_j(x_j^{(i)} \mid y)}{\delta(y \mid i)}$$

$$= \sum_{i=1}^{m} \sum_{y=1}^{k} \delta(y \mid i) \left[\log p(y) + \sum_{j=1}^{n} \log p_j(x_j^{(i)} \mid y) - \log \delta(y \mid i) \right]$$

Applying EM Algorithm to NB: M-Step (Contd')

$$\sum_{i=1}^{m} \sum_{y=1}^{k} \delta(y \mid i) \log \frac{p(x^{(i)}, z^{(i)} = y)}{\delta(y \mid i)}$$

$$= \sum_{i=1}^{m} \sum_{y=1}^{k} \delta(y \mid i) \left[\log p(y) + \sum_{j=1}^{n} \log p_j(x_j^{(i)} \mid y) - \log \delta(y \mid i) \right]$$

$$= \sum_{i=1}^{m} \sum_{y=1}^{k} \delta(y \mid i) \log p(y) + \sum_{i=1}^{m} \sum_{y=1}^{k} \sum_{j=1}^{n} \delta(y \mid i) \log p_j(x_j^{(i)} \mid y)$$

$$- \sum_{i=1}^{m} \sum_{y=1}^{k} \delta(y \mid i) \log \delta(y \mid i)$$

Applying EM Algorithm to NB: M-Step (Contd')

$$\sum_{i=1}^{m} \sum_{y=1}^{k} \sum_{j=1}^{n} \delta(y \mid i) \log p_j(x_j^{(i)} \mid y)$$

$$= \sum_{y=1}^{k} \sum_{j=1}^{n} \left(\sum_{i:x_j^{(i)}=0} \delta(y \mid i) \right) \log p_j(x = 0 \mid y)$$

$$+ \sum_{y=1}^{k} \sum_{j=1}^{n} \left(\sum_{i:x_j^{(i)}=1} \delta(y \mid i) \right) \log p_j(x = 1 \mid y)$$

$$= \sum_{y=1}^{k} \sum_{j=1}^{n} \sum_{x \in \{0,1\}} \left(\sum_{i:x_j^{(i)}=x} \delta(y \mid i) \right) \log p_j(x \mid y)$$

Applying EM Algorithm to NB: M-Step (Contd')

$$\max \quad \sum_{i=1}^{m}\sum_{y=1}^{k}\delta(y \mid i)\log p(y) + \sum_{y=1}^{k}\sum_{j=1}^{n}\sum_{x\in\{0,1\}}\left(\sum_{i:x_j^{(i)}=x}\delta(y \mid i)\right)\log p_j(x \mid y)$$

$$s.t. \quad \sum_{y=1}^{k} p(y) = 1$$

$$\sum_{x\in\{0,1\}} p_j(x \mid y) = 1, \ \forall y = 1,\cdots,k, \ \forall j = 1,\cdots,n$$

$$p(y) \geq 0, \ \forall y = 1,\cdots,k$$

$$p_j(x \mid y) \geq 0, \ \forall j = 1,\cdots,n, \ \ \forall x \in \{0,1\}, \ \ \forall y = 1,\cdots,k$$

Applying EM Algorithm to NB: M-Step (Contd')

- Problem I

$$\max \quad \sum_{i=1}^{m} \sum_{y=1}^{k} \delta(y \mid i) \log p(y)$$

$$s.t. \quad \sum_{y=1}^{k} p(y) = 1$$

$$p(y) \geq 0, \quad \forall y = 1, \cdots, k$$

- Solution (by Lagrange multiplier):

$$p(y) = \frac{\sum_{i=1}^{m} \delta(y \mid i)}{\sum_{i=1}^{m} \sum_{y=1}^{k} \delta(y \mid i)} = \frac{1}{m} \sum_{i=1}^{m} \delta(y \mid i)$$

Applying EM Algorithm to NB: M-Step (Contd')

- Problem II

$$
\max \sum_{y=1}^{k} \sum_{j=1}^{n} \sum_{x \in \{0,1\}} \left(\sum_{i:x_j^{(i)}=x} \delta(y \mid i) \right) \log p_j(x \mid y)
$$

$$
s.t. \quad \sum_{x \in \{0,1\}} p_j(x \mid y) = 1, \ \forall y = 1, \cdots, k, \ \forall j = 1, \cdots, n
$$

$$
p_j(x \mid y) \geq 0, \ \forall j = 1, \cdots, n, \ \forall x \in \{0,1\}, \ \forall y = 1, \cdots, k
$$

- Solution (by Lagrange multiplier):

$$
p_j(x \mid y) = \frac{\sum_{i:x^{(i)}=x} \delta(y \mid i)}{\sum_{x' \in \{0,1\}} \sum_{i:x_j^{(i)}=x'} \delta(y \mid i)} = \frac{\sum_{i:x^{(i)}=x} \delta(y \mid i)}{\sum_{i=1}^{m} \delta(y \mid i)}
$$

Thanks!

Q & A