CS789: Machine Learning and Neural Network

Bayesian learning

Jakramate Bootkrajang

Department of Computer Science
Chiang Mai University

What will we learn in this lecture?

- We will learn about
  - Bayes' rule
  - We will construct various classifiers using Bayes' rule

Bayes' Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- $P(Y)$: prior belief, prior probability, or simply prior.
  - Probability of observing class Y.
- $P(X|Y)$: likelihood
  - (Relative) probability of seeing X in class Y
- $P(X) = \sum_Y P(X|Y)P(Y)$: data evidence
- $P(Y|X)$: a posteriori probability
  - Probability of class Y after having seen the data X

A Side Note on Probability

- A likelihood function $L(\theta|X)$ is a function describing probability of a parameter given an outcome.
- For a dataset $S = (x_1, \ldots, x_m)$ the likelihood of parameter θ is given by

$$L(\theta|X) = P(X = x_1|\theta) \cdot P(X = x_2|\theta) \cdots P(X = x_m|\theta)$$

A Side Note on Probability

- Suppose we have two dices h_1 and h_2
  - Say, h_1 is fair but h_2 is biased
- The probability of getting i given the h_1 dice is called conditional probability, denoted by $P(i|h_1)$
- Pick a dice at random with $P(h_j) : j = 1, 2$. The probability for picking the h_j dice and getting an i with the dice is called joint probability, and is $P(i, h_j) = P(h_j)P(i|h_j)$
- The so-called marginal probability is $P(i) = \sum_{h_j} P(i, h_j)$.

Bayes' decision rule

Consider a binary classification: $Y \in \{-1, 1\}$, we can construct a classfier with minimal probability of error if we define

Definition (Bayes decision rule)

$$h^*(x) = \begin{cases} 1 & P(Y = 1|X = x) > 1/2 \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

Theorem

For any classifier $h : X \to \{-1, 1\}$,

$$P(h^*(X) \neq Y) \leq P(h(X) \neq Y), \quad (2)$$

that is, h^* *is the optimal classifier.*

Building a classifier using Bayes rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Goal:

- Calculate probabilities of how likely to see label Y when X is presented, $P(Y|X)$

How ?

- Find $P(X|Y)$
- Find $P(Y)$
- Find $P(X)$

Building a classifier using Bayes rule

- Way to obtain $P(X|Y)$
  - $P(X|Y)$ can be given.
  - $P(X|Y)$ can be modelled using discrete probability distribution.
    - ⋆ We can count number of time X occurs to estimate its likelihood given Y.
  - $P(X|Y)$ can be modelled using continuous probability distribution.
    - ⋆ We estimate parameters of the distribution.
- Way to obtain $P(Y)$, find ratio $\frac{\#Y}{N}$
- Way to obtain $P(X)$, find marginal probability $\sum_Y P(X|Y)P(Y)$

Two philosophies of estimating $P(Y|X)$

- Maximum Likelihood: assume equal priors
  - $h_{ML}(X) = \text{argmax}_y P(Y = y|X) = \text{argmax}_y \frac{P(X|Y=y)\times 0.5}{P(X)}$
  - Often used when we have very little idea about the data.
- Maximum a Posteriori: consider priors
  - $h_{MAP}(X) = \text{argmax}_y P(Y = y|X) = \text{argmax}_y \frac{P(X|Y=y)\times P(Y=y)}{P(X)}$
  - Generally gives better performance if we have the priors.

A word about the Bayesian Framework

- Allows us to combine observed data and prior knowledge
- Provides practical learning algorithms
- It is a generative approach, which offers a useful conceptual framework
  - This means that any kind of objects (e.g. time-series, trees, etc.) can be classified, based on a probabilistic model specification

Case 1: $P(X|Y)$ is given

Does a patient have cancer or not?
A patient takes a lab test and the result comes back positive. It is known that the test returns a correct positive result in only 98% of the cases and a correct negative result in only 97% of the cases. Furthermore, only 0.008 of the entire population has this disease.

Working out the variables

- $Y = \{cancer, \neg cancer\}$, $X = \{positive, negative\}$
- To decide whether the patient has cancer we have to calculate
  - The posterior probability the the patient has cancer, $P(Y = cancer|X = positive)$
  - The posterior probability the the patient does not have cancer, $P(Y = \neg cancer|X = positive)$
- According to the Bayes decision rule, we pick y which gives $P(Y = y|X = x) > 0.5$

ML Solution

1. The posterior probability of having cancer.

$$P(Y = cancer|X = positive) = \frac{P(X = positive|Y = cancer) \times 0.5}{P(X = positive)}$$
$$P(X = positive) = P(X = positive|Y = cancer)P(cancer) + P(X = positive|Y = \neg cancer)P(\neg cancer)$$
$$= \ldots\ldots\ldots$$

2. The posterior probability of being healthy.

$$P(Y = \neg cancer|X = positive) = \frac{P(X = positive|Y = \neg cancer) \times 0.5}{P(X = positive)}$$
$$= \ldots\ldots\ldots$$

3. Diagnosis ??

MAP Solution

1. The posterior probability of having cancer.

$$P(Y = cancer|X = positive) = \frac{P(X = positive|Y = cancer) \times 0.008}{P(X = positive)}$$
$$P(X = positive) = P(X = positive|Y = cancer)P(cancer) + P(X = positive|Y = \neg cancer)P(\neg cancer)$$
$$= \ldots\ldots\ldots$$

2. The posterior probability of being healthy.

$$P(Y = \neg cancer|X = positive) = \frac{P(X = positive|Y = \neg cancer) \times 0.992}{P(X = positive)}$$
$$= \ldots\ldots\ldots$$

3. Diagnosis ??

Case 2: Discrete $P(X|Y)$

The dataset: (W,F), (BR,F), (W,A), (B,F), (B,F), (BR,F), (W,A)

- Assume we have a set of data which classifies dog friendliness based on its colour.
- $Y = \{Aggressive, Friendly\}$, $X = \{White, BRown, Black\}$
- If we see new white dog would it be friendly ?

ML Solution

1. The posterior probability of being friendly.

$$P(Y = friendly|X = white) = \frac{P(X = white|Y = friendly) \times 0.5}{P(X = white)}$$
$$P(X = white) = P(X = white|Y = friendly)P(Y = friendly) + P(X = white|Y = aggressive)P(Y = aggressive)$$
$$= \ldots\ldots\ldots$$

2. The posterior probability of being aggressive.

$$P(Y = aggressive|X = white) = \frac{P(X = white|Y = aggressive) \times 0.5}{P(X = white)}$$
$$= \ldots\ldots\ldots$$

3. Diagnosis ??

The Naive Bayes Classifier (1/2)

- What if our example has several attributes? $x = \{a_1, a_2, \ldots, a_n\}$
- The problem is $P(X, Y) = P(Y)P(X|Y)$ factorised into a long sequence.
- By chain rule,

$$\begin{aligned} P(X, Y) &= P(Y)P(a_1, \ldots, a_m|Y) \\ &= P(Y)P(a_1|Y)P(a_2, \ldots, a_m|Y, a_1) \\ &= P(Y)P(a_1|Y)P(a_2|Y, a_1)P(a_3, \ldots, a_m|Y, a_1, a_2) \end{aligned}$$

- The naive assumption assumes that each feature a_i is conditionally independent of every other feature a_j for $j \neq i$

The Naive Bayes Classifier (2/2)

- So we have $P(a_i|Y, a_j, \ldots) = P(a_i|Y)$ and so on.
- Which gives: $P(X|Y) = P(a_1, \ldots, a_m|Y) = \prod_i P(a_i|Y)$
- A Bayesian classifier that uses the Naive assumption is called The Naive Bayes classifier.
- One of the most practical methods widely used in,
  - Medical applications.
  - Text classification.

Example of Naive classifier: Playing Tennis

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	No
Day2	Sunny	Hot	High	Strong	No
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	No
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	No
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	No

Naive Bayes Solution

- Classify any new data point $x = (a_1, \ldots, a_m)$ as
- $h_{naive}(X) = \text{argmax}_Y \, P(Y)P(X|Y) = \text{argmax}_Y \, P(Y) \prod_i P(a_i|Y)$
- We need to estimate the parameters from the training examples
  - For each class y: $\hat{P}(Y = y)$
  - For each feature a_i: $\hat{P}(a_i|Y)$
- Based on the examples in the table, classify the following x
- $x = \{sunny, cool, high, strong\}$, Play tennis or not ?

The working

$$
\begin{aligned}
h_{naive} &= \text{argmax}_{y \in [yes,no]} P(Y=y) P(X=x|Y=y) \\
&= \text{argmax}_{y \in [yes,no]} P(Y=y) \prod_i P(a_i|Y=y) \\
&= \text{argmax}_{y \in [yes,no]} P(Y=y) P(sunny|Y=y) P(cool|Y=y) \\
&\quad P(high|Y=y) P(strong|Y=y)
\end{aligned}
$$

- Now find
  - $P(Y=yes) = 9/14 = 0.64$
  - find $P(sunny|Y=yes) = 2/9 = 0.22$
  - find $P(cool|Y=yes) = 3/9 = 0.33$
  - find $P(high|Y=yes) = \ldots\ldots$
  - find $P(strong|Y=yes) = \ldots\ldots$ and so on....

Exercise

- Assume we have a data set described the following three variables:
  Hair = B,D, where B=blonde, D=dark.
  Height = T,S, where T=tall, S=short.
  Country = G,P, where G=Greenland, P=Poland.
- You are given the following training data set (Hair, Height, Country):
  (B,T,G), (D,T,G), (D,T,G), (D,T,G), (B,T,G), (B,S,G), (B,S,G), (D,S,G), (B,T,G), (D,T,G), (D,T,G), (D,T,G), (B,T,G), (B,S,G), (B,S,G), (D,S,G), (B,T,P), (B,T,P), (B,T,P), (D,T,P), (D,T,P), (D,S,P), (B,S,P), (D,S,P).
- Now, suppose you observe a new individual tall with blond hair, and you want to use these training data to determine the most likely country of origin.
- Compute the maximum a posteriori (MAP) answer to the above question, using the Naive Bayes assumption.

Learning to classify text

- The attributes (features) are the words
- NB classifiers are one of the most effective for this task

Representation of text: bag of words

> I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.

Representation of text: bag of words

Predefine **vocabolary set** V and highlight $w \in V$.

> I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.

Representation of text: bag of words

great	2
love	2
recommend	1
laugh	1
terrible	0
happy	1
sad	0
⋮	⋮

Parameter estimation

- Simply use the frequencies in the data (Case 2)
- Class prior probability:

$$P(Y = y_j) = \frac{\text{count}(Y=y_j)}{m}$$

- Likelihood

$$P(w_i|Y = y_j) = \frac{\text{count}(w_i, Y=y_j)}{\sum_{w \in V} \text{count}(w, Y=y_j)}$$

- Problem of the above is if no training documents contain the word **fantastic** in class **positive**, then $P(\text{“}fantastic\text{”}|positive) = 0$
- So $P(positive|X_{new})$ will always be zero.

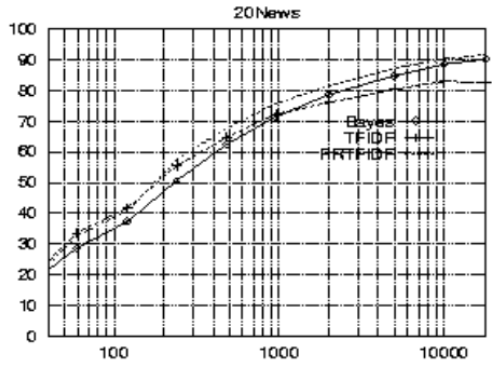
Laplace smoothing for Naive Bayes

To pretend that you have seen each of all the words in V at least α times.

$$P(w_i|Y) = \frac{\text{count}(w_i, Y) + \alpha}{\sum_{w \in V} (\text{count}(w, Y) + \alpha)}$$
$$= \frac{\text{count}(w_i, Y) + \alpha}{(\sum_{w \in V} \text{count}(w, Y)) + \alpha|V|}$$

Here, α is called smoothing parameter (aka *hyper-parameter*) which often be tuned using cross-validation.

Example of 20-Newsgroups text classification using NB



Accuracy vs. Training set size (1/3 withheld for test)

Summary

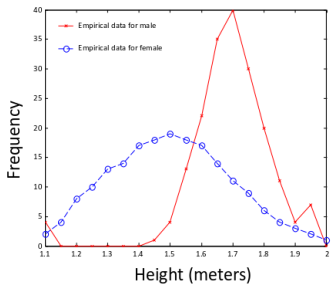
- Bayes' rule can be turned into a classifier
- Maximum A Posteriori (MAP) hypothesis estimation incorporates prior knowledge; Maximum Likelihood doesn't
- Naive Bayes classifier is a simple but effective Bayesian classifier for vector data (i.e. data with several attributes) that assumes that attr. are independent given the class.
- Bayesian classification is a generative approach to classification.

Case 3: Motivations

- We have already seen how Bayes rule can be turned into a classifier.
- Our examples so far only consider discrete attributes.
  - E.g. {sunny, warm}, {positive,negative}
- Today we learn how to do this when the data attributes are continuous.

An example problem

- Task: predict gender of individuals based on their heights.
- Given
  - 100 height examples of women.
  - 100 height examples of man.
- Encode class label of male as $y = 1$ and female as $y = 0$. So, $y \in \{0, 1\}$.



Class posterior

- From Bayes rule we can obtain the class posteriors of male:

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

- The denominator is the probability of measuring the height x irrespective of the class.

Modelling $p(x|y)$ using continuous probability distribution

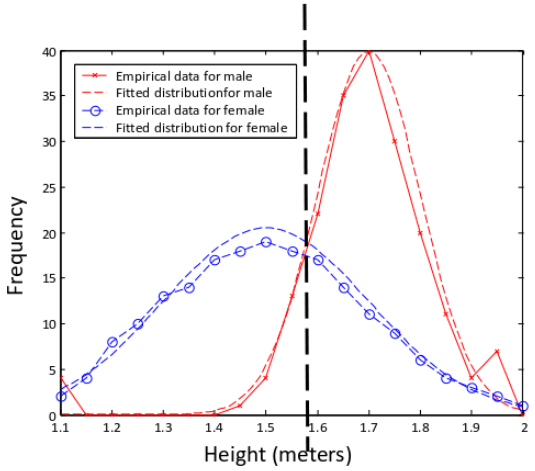
- Our measurements are heights. This is our data, x.
- Class-conditional likelihoods
  - $p(x|y=1)$: probability that a male has height x metres.
  - $p(x|y=0)$: probability that a female has height x metres.
- We will model each class by a Gaussian distribution. (Other distribution is possible)

Univariate Gaussian (Normal Distribution)

$$p(x|y=k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\{-\frac{(x-\mu_k)^2}{2\sigma_k^2}\}$$

- where μ_k is the mean(centre), and σ_k^2 is the variance (spread). These are parameters that describe the distributions.
- We will have separate Gaussian for each class. So, the female class will have μ_0 as its mean, and σ_0^2 as its variance. And male class with m_1 and σ_1^2.
- We will estimate these parameters from the data.

Illustration - our 1D example



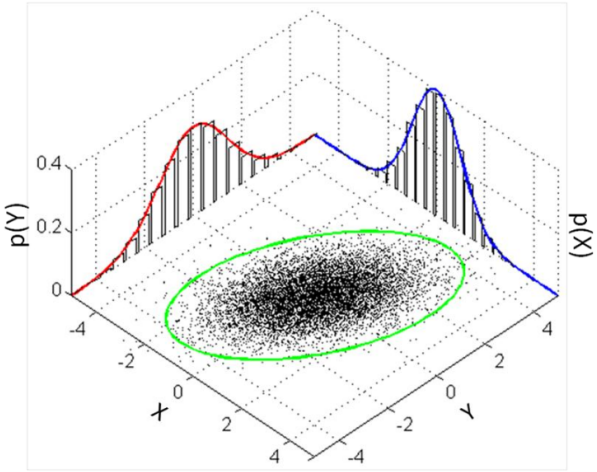
Multivariate Gaussian

- Let $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_d\}$. Let $k \in \{0, 1\}$

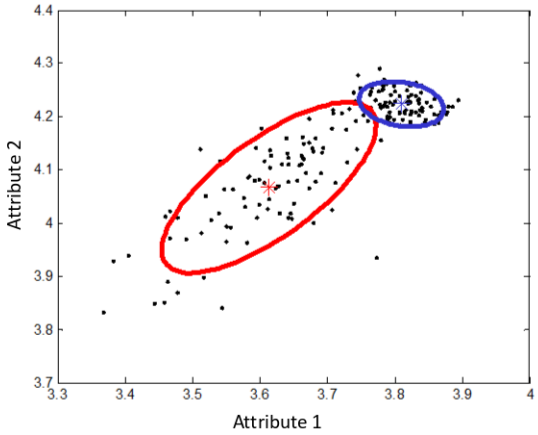
$$p(\mathbf{x}|y = k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \mu_k)^t \Sigma_k^{-1}(\mathbf{x} - \mu_k)\}$$

- where μ_k is the mean vector, and Σ_k is the covariance matrix.
- These parameters get estimated from the data.

Illustration - 2D example



Example with 2 classes



Class priors

- Class prior: the probability of seeing male example (and female example).
- Since in this example we had the same number of males and females, we *empirically* calculate,

$$p(y = 1) = p(y = 0) = \frac{100}{100 + 100} = 0.5$$

- These are priors of class membership and they could be set before measuring any data.
- The class prior can be useful in cases where class proportions are imbalanced.

Discriminant function

- According to Bayes decision rule, we will predict $y = 1$ if $p(y = 1|x) > 1/2$ and $y = 0$ otherwise.
- We can formulate the above rule as a mathematical function.

$$f_1(x) = \mathbb{1}\Big(\frac{p(y=1|x)}{p(y=0|x)} > 1\Big)$$

- Or equivalently

$$f_2(x) = \mathbb{1}\Big(\log \frac{p(y=1|x)}{p(y=0|x)} > 0\Big)$$

The sign of f_2 defines the prediction $f_2(x) > 0 =$ male, $f_2(x) \leq 0 =$ female

- Such functions are called discriminant functions.

Discriminant Analysis

- Recall our discriminant function $f_2(x) = \log \frac{p(y=1|x)}{p(y=0|x)}$
- We'd like to know what decision boundary a particular Σ will induced.
- We write (for normal density and $\omega_i \overset{\text{def}}{=} p(y = k)$)

$$\begin{aligned}
f_2(x) &= \log \frac{p(x|\mu_1, \Sigma_1)\omega_1}{p(x|\mu_0, \Sigma_0)\omega_0} \\
&= \log p(x|\mu_1, \Sigma_1) + \log \omega_1 - \log p(x|\mu_0, \Sigma_0) - \log \omega_0 \\
&= \dots \\
&= -\frac{1}{2}(x-\mu_1)^t \Sigma_1^{-1}(x-\mu_1) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_1| + \log \omega_1 \\
&\quad + \frac{1}{2}(x-\mu_0)^t \Sigma_0^{-1}(x-\mu_0) + \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma_0| - \log \omega_0
\end{aligned}$$

Discriminant Analysis: case $\Sigma_1 = \Sigma_0 = \sigma^2 I$

- The determinant is $|\Sigma| = \sigma^{2d}$
- And $\Sigma^{-1} = (1/\sigma^2)I$

$$\begin{aligned}
f_2(x) &= -\frac{1}{2}(x-\mu_1)^t \frac{I}{\sigma^2}(x-\mu_1) + \log \omega_1 \\
&\quad + \frac{1}{2}(x-\mu_0)^t \frac{I}{\sigma^2}(x-\mu_0) - \log \omega_0 \\
&= -\frac{1}{2\sigma^2}(x^t x - 2\mu_1^t x + \mu_1^t \mu_1) + \log \omega_1 \\
&\quad + \frac{1}{2\sigma^2}(x^t x - 2\mu_0^t x + \mu_0^t \mu_0) - \log \omega_0 \\
&= -\frac{1}{\sigma^2}(\mu_1^t x - \frac{1}{2}\mu_1^t \mu_1) + \frac{1}{\sigma^2}(\mu_0^t x - \frac{1}{2}\mu_0^t \mu_0) + \log \frac{\omega_1}{\omega_0}
\end{aligned}$$

Discriminant Analysis: case $\Sigma_1 = \Sigma_0 = \sigma^2 I$

$$\begin{aligned}
-\frac{1}{\sigma^2}(\mu_1^t x - \frac{1}{2}\mu_1^t \mu_1) + \frac{1}{\sigma^2}(\mu_0^t x - \frac{1}{2}\mu_0^t \mu_0) + \log \frac{\omega_1}{\omega_0} &= 0 \\
-(\mu_1^t x - \frac{1}{2}\mu_1^t \mu_1) + (\mu_0^t x - \frac{1}{2}\mu_0^t \mu_0) + \sigma^2 \log \frac{\omega_1}{\omega_0} &= 0 \\
(\mu_0 - \mu_1)^t x + \frac{1}{2}\mu_1^t \mu_1 - \frac{1}{2}\mu_0^t \mu_0 + \sigma^2 \log \frac{\omega_1}{\omega_0} &= 0 \\
(\mu_0 - \mu_1)^t x + \frac{1}{2}(\mu_1^t \mu_1 - \mu_0^t \mu_0) + \sigma^2 \log \frac{\omega_1}{\omega_0} &= 0 \\
(\mu_0 - \mu_1)^t x + \frac{1}{2}(\mu_1 - \mu_0)^t(\mu_1 + \mu_0) + \sigma^2 \log \frac{\omega_1}{\omega_0} &= 0 \\
(\mu_0 - \mu_1)^t x - (\mu_0 - \mu_1)^t \Big[\frac{1}{2}(\mu_1 + \mu_0) + \frac{\sigma^2}{(\mu_0 - \mu_1)} \log \frac{\omega_1}{\omega_0}\Big] &= 0 \\
w^t(x - x_0) &= 0
\end{aligned}$$

Discriminant Analysis: case $\Sigma_1 = \Sigma_0 = \sigma^2 I$

$$w^t(x - x_0) = 0$$

- This defines a hyperplane through point x_0 and orthogonal to w.
- Since $w = (\mu_0 - \mu_1)$ the hyperplane is a plane normal to the line linking the means.
- The plane cut the line at x_0.
- If $\omega_1 = \omega_0$ then $x_0 = (\mu_0 - \mu_1)/2$, the midpoint between the means.
- In other cases, x_0 shifts away from the more likely mean (from class with larger ω or larger prior)

Discriminant Analysis: case $\Sigma_1 = \Sigma_0 = \Sigma$

- Along the same line of analysis we found that in this case

$$w = \Sigma^{-1}(\mu_i - \mu_j)$$
$$x_0 = \frac{1}{2}(\mu_i - \mu_j) - \frac{\log[p(\omega_i)/p(\omega_j)]}{(\mu_i - \mu_j)^t \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$

- The decision boundary is still linear and ω controls the intercept on the line linking the means.
- However, the hyperplane is not orthogonal to the line between the means due to the covariance, $w = \Sigma^{-1}(\mu_i - \mu_j)$

Discriminant Analysis: case $\Sigma_1 \neq \Sigma_0 =$ arbitrary

- In the last case we found that

$$f_2(x) = -\frac{1}{2}(x - \mu_1)^t \Sigma_1^{-1}(x - \mu_1) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_1| + \log\omega_1$$
$$+ \frac{1}{2}(x - \mu_0)^t \Sigma_0^{-1}(x - \mu_0) + \frac{d}{2}\log 2\pi + \frac{1}{2}\log|\Sigma_0| - \log\omega_0$$

- The decision boundary is quadratic, since things cannot be simplified.
- The non-linearity of this form leads to more powerful classifier for tackling data which is not linearly-separable.

Gaussian's parameters estimation

The covariance

$$\Sigma_k = \frac{\sum_{i=1}^{m_k}(x_i - \mu_k)(x_i - \mu_k)^t}{m_k}$$

The mean

$$\mu_k = \frac{\sum_{i=1}^{m_k} x_i}{m_k}$$

The prior

$$\omega_k = p(y = k) = \frac{m_k}{m}$$

Naive assumption

- The full covariance are $d \times d$
- In many situation there is not enough data to estimate full covariance.
- The Naive Bayes is again useful and tends to work well in practice.
- Using Naive assumption the covariance becomes diagonal.

Multi-class classification

- We may have more than two classes. Say, 'Healthy', 'Disease 1', 'Disease 2'.
- Our Gaussian classifier is easy to use in multi-class problem.
- We compute posterior probability for each of the classes.
- We predict class with highest posterior.

Summary

- This type of classifier is call *Generative* because it makes an assumption that the points in each class are generated by some distribution i.e., Gaussian distribution in our example.
- One can model the discriminant function directly. That is called *Discriminative* classifier – (next week)

References

- Machine learning course by Ata Kabán. `http://www.cs.bham.ac.uk/~axk/ML_new.htm`