K-Nearest Neighbor Classifiers (KNNC)

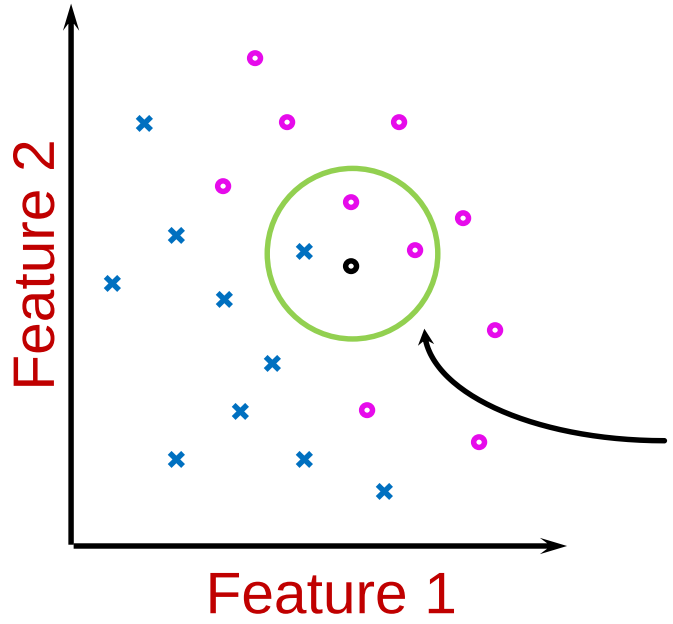
J.-S. Roger Jang (張智星)
jang@mirlab.org
http://mirlab.org/jang
MIR Lab, CSIE Dept.
National Taiwan University

Concept of KNNC

- Concept: 近朱者赤、近墨者黑
- Two Steps:

Quiz!

  - Find the first k nearest neighbors of a given point.
  - Determine the class of the given point by voting among k nearest neighbors.

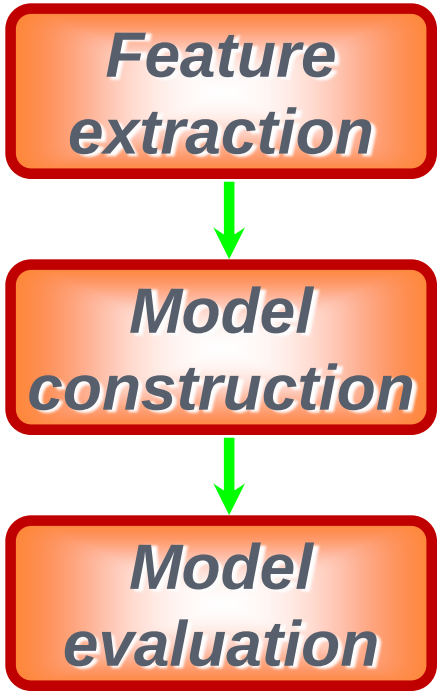


× : class-A point
∘ : class-B point
⚬ : point with unknown class

3 nearest neighbors
➔ The point is classified as B via 3NNC.

Flowchart for KNNC

General flowchart of PR:



KNNC:

From raw data to features

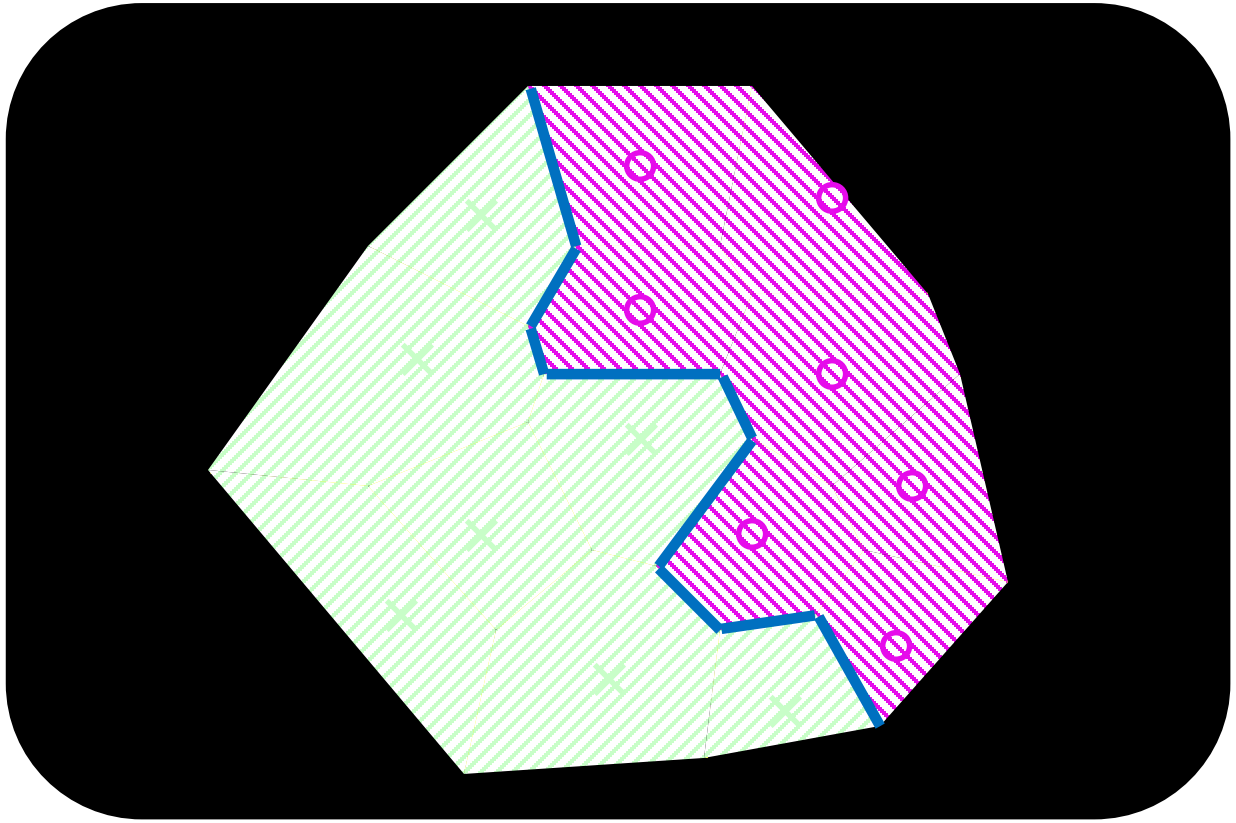
Clustering (optional)

KNNC evaluation on test dataset

Decision Boundary for 1NNC

- Voronoi diagram: piecewise linear boundary

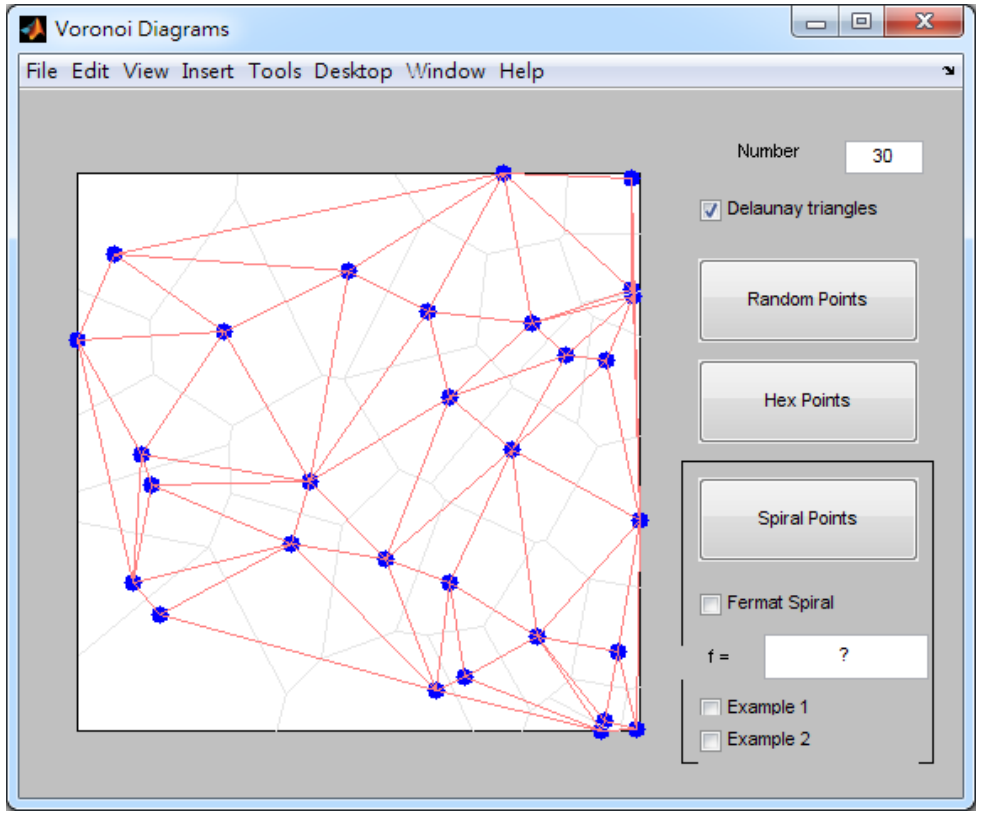
Quiz!



More about Voronoi diagrams

Demos by Cleve Moler

- Cleve's Demos of Delaunay triangles and Voronoi diagram
  - books/dcpr/example/cleve/vshow.m



Natural Examples of Voronoi Diagrams (1/2)

Characteristics of KNNC

- Strengths of KNNC
  - Intuitive
  - No computation for model construction
- Weakness of KNNC
  - Massive computation required when dataset is big
  - No straightforward way
    - To determine the value of K
    - To rescale the dataset along each dimension

Quiz!

Preprocessing of Feature Normalization

- Z normalization or z score
  - To have zero mean and unit variance along each feature
- Min-max normalization
  - To have a specific range, such as [0, 1], along each feature

Quiz!

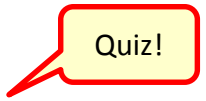
Let $\mathbf{x} = [x_1, x_2, \cdots, x_n]$ be the values of a specific feature of a dataset

Z normalization : $\hat{x}_i = \dfrac{x_i - \mu}{\sigma}$, with μ and σ^2 being the sample mean and sample variance of $\mathbf{x}$ respectively

Min - max normalization : $\hat{x}_i = \dfrac{x_i - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}$ to have a range of $[0,1]$

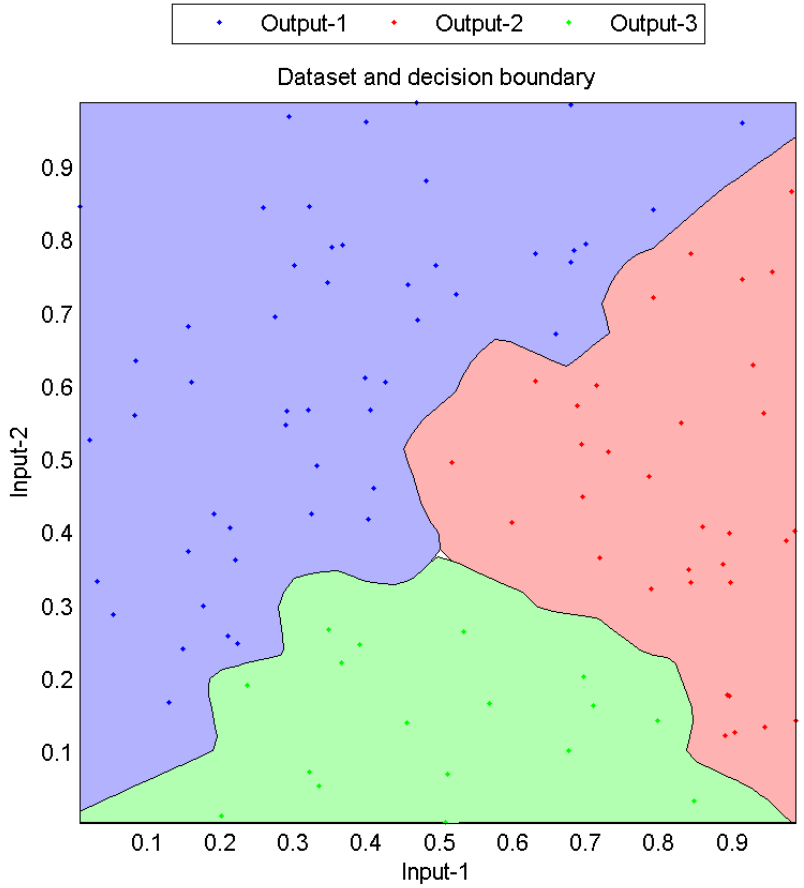
Variants for KNNC

- Many variants of KNNC:
  - Nearest prototype classification
    - Single prototype for each class ➔ Use “mean” or “average”
    - Several prototypes for each class ➔ Use “k-means clustering”
  - Distance-weighted votes
  - Edited nearest neighbor classification
  - k+k-nearest neighbor

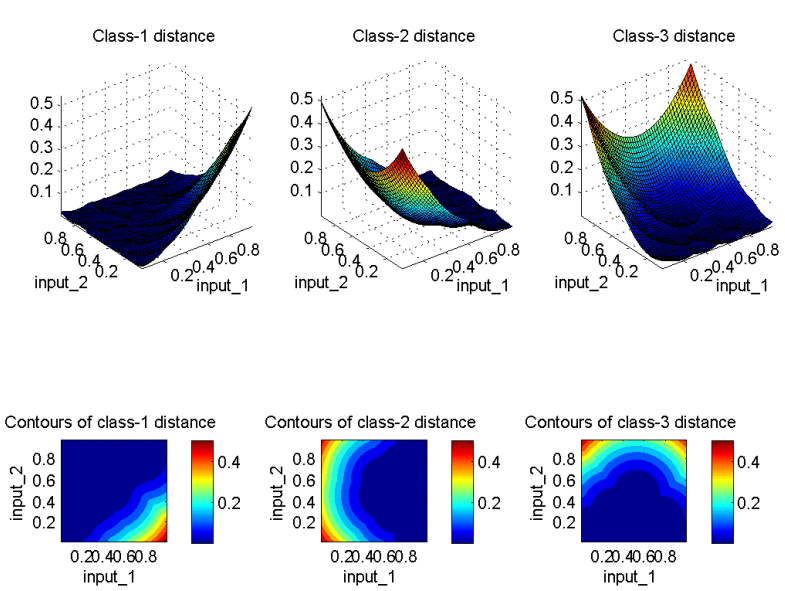


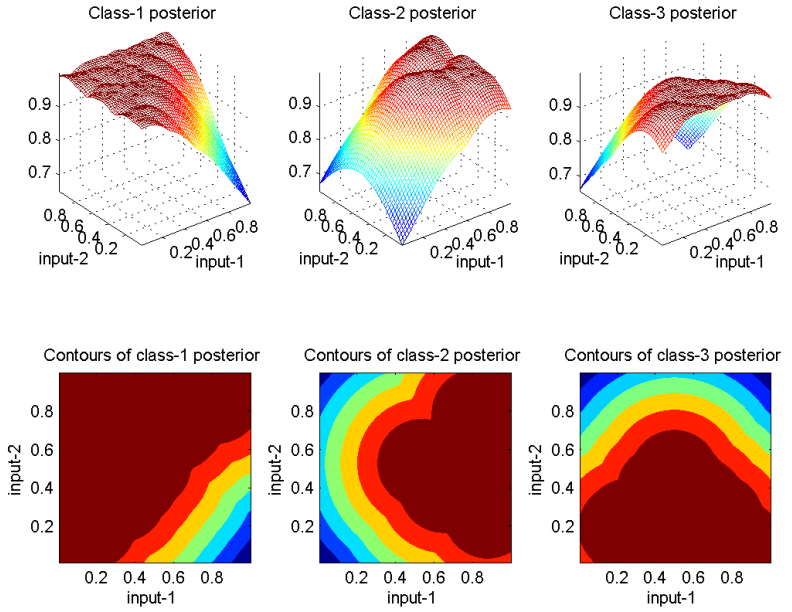
1NNC Decision Boundaries

- 1NNC Decision boundaries



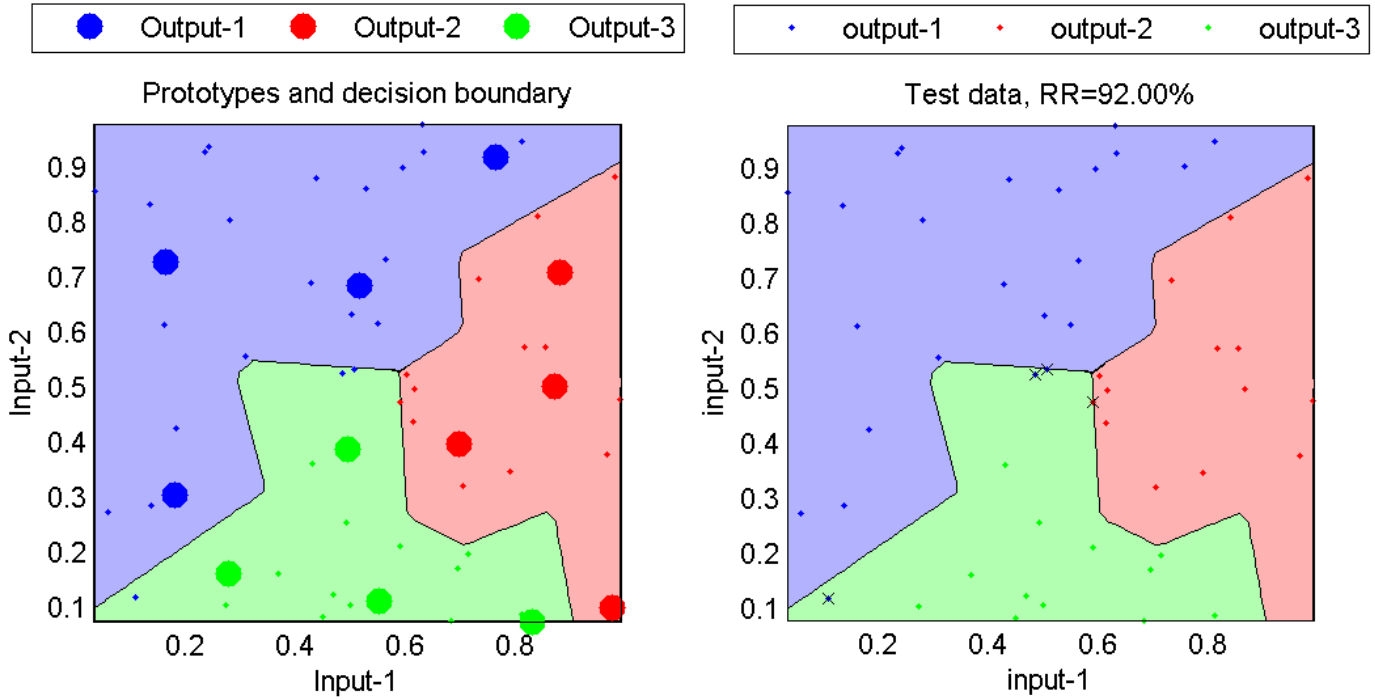
1NNC Distance/Posterior as Surfaces and Contours





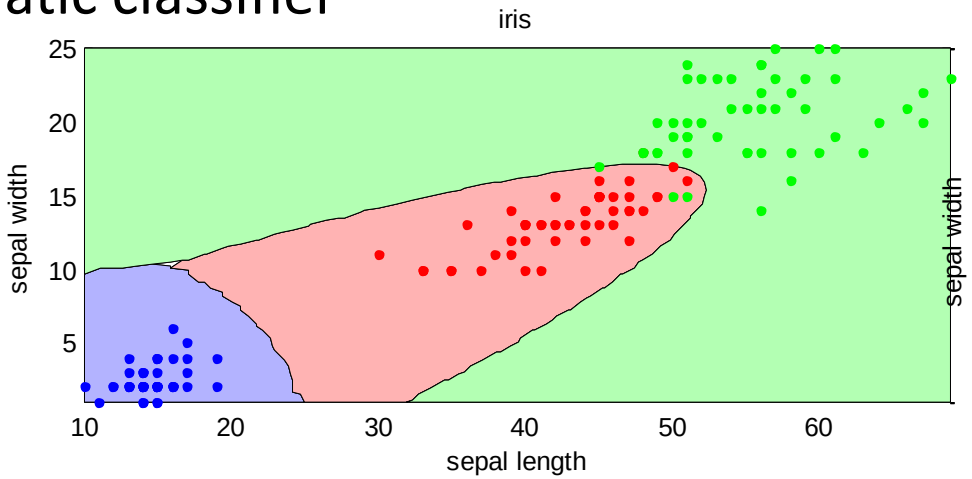
Using Prototypes in KNNC

- No. of prototypes for each class is 4.

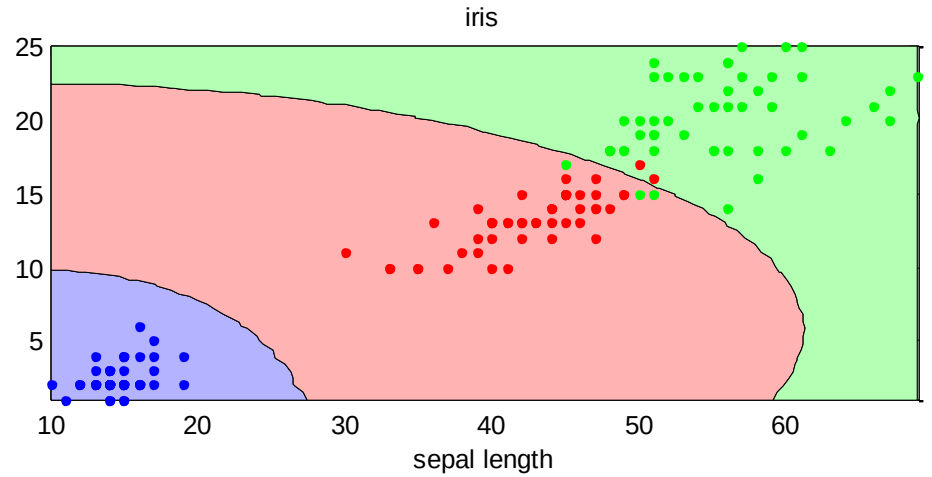


Decision Boundaries of Different Classifiers

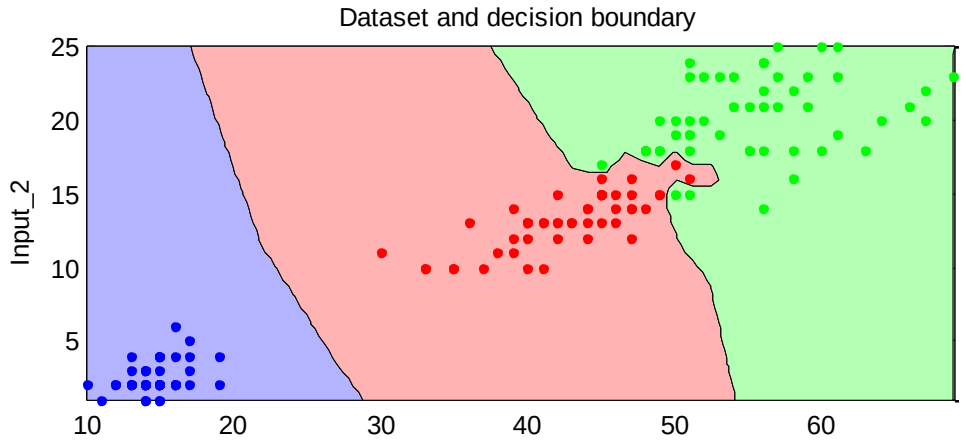
- Quadratic classifier



- Naive Bayes classifier



- 1NNC classifier



Maximum Likelihood Estimate

Jyh-Shing Roger Jang (張智星)
CSIE Dept, National Taiwan University

2018/10/15

Intro. to Maximum Likelihood Estimate

- MLE
  - Maximum likelihood estimate
- Goal:
  - Given a dataset with no labels, how can we find the best statistical model with the optimum parameters to describe the data?
- Applications
  - Prediction
  - Analysis

What Are Statistical Models?

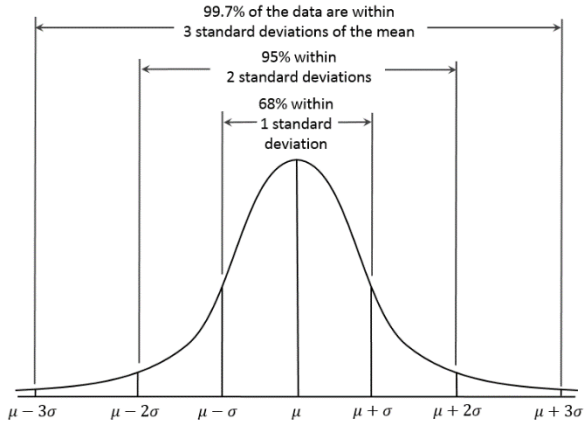
- Statistical models are used to describe the probabilities of random variables
  - Discrete variables ➔ Probability functions
  - Continuous variables ➔ Probability density functions (PDF)
- Examples
  - Discrete variables
    - The outcome of tossing a coin or a die

  - Continuous variables
    - The distance to the bull eye when throwing a dart
    - The time needed to run 100-m dash
    - The heights of second-grade students

Personalized PDF!

More about Models

- Discrete variables
  - Outcome of tossing a coin ➔ Pr{head}=1/2, Pr{tail}=1/2
- Continuous variables
  - Distance to the bull's eye when throwing a dart ➔ A PDF of Gaussian or normal distribution



$$g\left(x;\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Quiz!

$$\Pr\{x\in[4,6]\}=\int_4^6 g(x;\mu,\sigma^2)dx$$

Probability of x in [4, 6]

Basic Steps in MLE

- Steps
  1. Perform a certain experiment to collect the data.
  2. Choose a parametric model of the data, with certain modifiable parameters.
  3. Formulate the likelihood as an objective function to be maximized.
  4. Maximize the objective function and derive the parameters of the model.
- Examples
  - Flip a coin ➔ To find the probabilities of head and tail
  - Throw a dart ➔ To find your PDF of distance to the bull eye

Probability Functions for Discrete Variables

- Flip an unfair coin 5 times to get 3 heads and 2 tails
  - By intuition: Pr{head}=3/5, Pr{tail}=2/5
  - By MLE
    - Assume these 5 tosses are independent events to have the overall probability

$$J(p,q) = p^3 q^2, \text{ with } p+q=1, p \geq 0, q \geq 0$$

$$\Rightarrow J(p) = p^3(1-p)^2$$

$$\Rightarrow \frac{dJ(p)}{dp} = 0$$

$$\Rightarrow p = 3/5, q = 2/5$$

Inequality of Arithmetic and Geometric Means

- AM-GM inequality

$$\frac{\sum_{i=1}^{n} x_i}{n} \geq \left(\prod_{i=1}^{n} x_i \right)^{1/n}, \text{ with } x_i \geq 0, \forall i$$

Quiz!

$$\text{The equality holds only when } x_1 = x_2 = \cdots = x_n.$$

- Proof of this inequality
  - Wikipedia
- How to use the inequality to solve MLE problem?

$$\frac{\frac{p}{3}+\frac{p}{3}+\frac{p}{3}+\frac{q}{2}+\frac{q}{2}}{5} \geq \left(\left(\frac{p}{3}\right)^3 \left(\frac{q}{2}\right)^2 \right)^{1/5}$$

$$\Rightarrow p^3q^2 \text{ achieves its maximum when } \frac{p}{3} = \frac{q}{2} \Rightarrow p = \frac{3}{5}, q = \frac{2}{5}$$

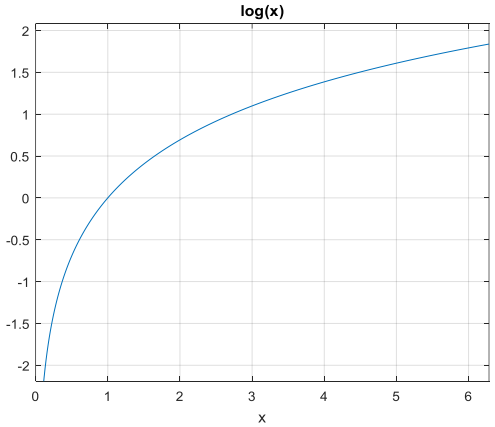
How to Prove AM-GM Inequality?

- Jensen's inequality

How to prove it?

$y = \ln x$ is a concave function $\Rightarrow$

$$\ln\left(\frac{mp+nq}{m+n}\right) \geq \frac{m\ln p + n\ln q}{m+n}, \text{ with } p,q,m,n>0$$



- Proof by induction

$$\ln\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) \geq \left(\frac{\sum_{i=1}^{n} \ln x_i}{n}\right), \text{ with } x_i > 0, \forall i$$

Proof by Induction

$$n=1 \Rightarrow x_1 \geq x_1$$

$$n=2 \Rightarrow \ln\left(\frac{x_1+x_2}{2}\right) \geq \frac{\ln x_1+\ln x_2}{2}. \text{(Or you can start with} \left(\sqrt{x_1}-\sqrt{x_2}\right)\geq 0)$$

$$n=3 \Rightarrow \ln\left(\frac{x_1+x_2+x_3}{3}\right)=\ln\left(\frac{2\left(\frac{x_1+x_2}{2}\right)+x_3}{3}\right) \geq \frac{2\ln\left(\frac{x_1+x_2}{2}\right)+\ln x_3}{3} \geq \frac{\ln x_1+\ln x_2+\ln x_3}{3}$$

$$n=k \text{ holds by assumption} \Rightarrow \ln\left(\frac{\sum_{i=1}^{k} x_i}{k}\right) \geq \left(\frac{\sum_{i=1}^{k} \ln x_i}{k}\right)$$

$$n=k+1 \Rightarrow \ln\left(\frac{\sum_{i=1}^{k} x_i + x_{k+1}}{k+1}\right)=\ln\left(\frac{k\frac{\sum_{i=1}^{k} x_i}{k}+x_{k+1}}{k+1}\right) \geq \frac{k\ln\left(\frac{\sum_{i=1}^{k} x_i}{k}\right)+\ln x_{k+1}}{k+1} \geq \frac{k\left(\frac{\sum_{i=1}^{k} \ln x_i}{k}\right)+\ln x_{k+1}}{k+1} = \frac{\sum_{i=1}^{k+1} \ln x_i}{k+1}$$

Probability Functions for Discrete Variables

- Toss a 3-side die for many times and obtain n_1 of side 1, n_2 of side 2, and n_3 of side 3, then what is the most likely probabilities for sides 1, 2, and 3, respectively?
  - Our intuition…
  - By MLE…

$$J(p,q,r) = p^{n_1} q^{n_2} r^{n_3}, \text{ with } p+q+r=1, p \geq 0, q \geq 0, r \geq 0$$

Quiz!

MLE for PDF of Continuous Variables of 1D

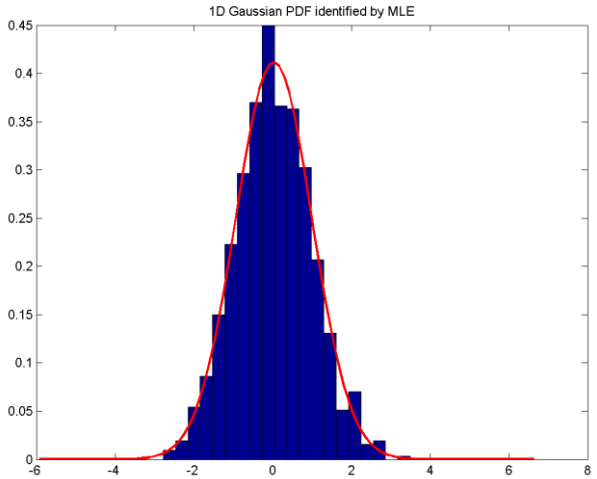
- Detailed coverage

$$g\left(x;\mu,\sigma^2\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

PDF

$$X=\{x_1,x_2\cdots x_n\}$$
$$\Rightarrow p\left(X;\mu,\sigma^2\right)=\prod_{i=1}^{n}g\left(x_i;\mu,\sigma^2\right)$$

Overall PDF, or likelihood



$$\begin{aligned}J(\mu,\sigma^2)&=\ln p\left(X;\mu,\sigma^2\right)\\&=\ln\left[\prod_{i=1}^{n}g\left(x_i;\mu,\sigma^2\right)\right]\\&=\sum_{i=1}^{n}\ln g\left(x_i;\mu,\sigma^2\right)\\&=\sum_{i=1}^{n}\left[-\frac{1}{2}\ln(2\pi)-\ln\sigma-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]\\&=-\frac{n}{2}\ln(2\pi)-n\ln\sigma-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2\end{aligned}$$

Log likelihood

MLE!

$$\frac{\partial J(\mu,\sigma^2)}{\partial\mu}=\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)=0\Rightarrow\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$$
$$\frac{\partial J(\mu,\sigma^2)}{\partial\sigma}=-\frac{n}{\sigma}-\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)\left(-\frac{x_i-\mu}{\sigma^2}\right)=0\Rightarrow\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\hat{\mu}\right)^2$$

Quiz!

MLE for PDF of Continuous Variables of ND

- Detailed coverage

$$g(x;\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

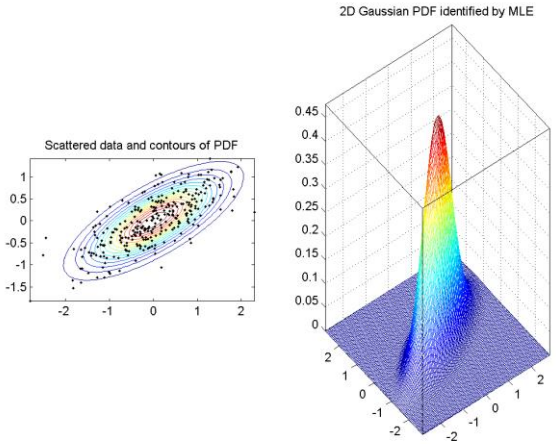
PDF

$$X=\{x_1,x_2\cdots x_n\}$$
$$\Rightarrow p(X;\mu,\Sigma)=\prod_{i=1}^{n}g(x_i;\mu,\Sigma)$$

Overall PDF, or likelihood

Log likelihood

$$\begin{aligned}J(\mu,\Sigma)&=\ln p(X;\mu,\Sigma)\\&=\ln\left[\prod_{i=1}^{n}g(x_i;\mu,\Sigma)\right]\\&=\sum_{i=1}^{n}\ln g(x_i;\mu,\Sigma)\\&=\sum_{i=1}^{n}\left[-\frac{d}{2}\ln(2\pi)-\frac{1}{2}\ln|\Sigma|-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]\\&=-\frac{nd}{2}\ln(2\pi)-\frac{n}{2}\ln|\Sigma|-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]\end{aligned}$$



$$\begin{aligned}\nabla_\mu J(\mu,\Sigma)&=-\frac{1}{2}\sum_{i=1}^{n}\left[-2\Sigma^{-1}(x_i-\mu)\right]\\&=\Sigma^{-1}\left(\sum_{i=1}^{n}x_i-n\mu\right)\end{aligned}$$
$$\nabla_\mu J(\mu,\Sigma)=0\Rightarrow\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\hat{\mu}\right)\left(x_i-\hat{\mu}\right)^T$$

MLE!

Q & A

- Questions
  - Can we choose other PDFs instead of Gaussian/normal distributions? ➔ Yes!
  - What are the other available PDFs?
  - How do I know the selected PDF is appropriate?

Quadratic Classifiers (QC)

J.-S. Roger Jang (張智星)
jang@mirlab.org
http://mirlab.org/jang
MIR Lab, CSIE Dept.
National Taiwan University

Review: PDF Modeling

- Goal:
  - Find a PDF (probability density function) that can best describe a given dataset
- Steps:
  - Select a class of parameterized PDF
  - Identify the parameters via MLE (maximum likelihood estimate) based on a given set of sample data
- Commonly used PDFs:
  - Multi-dimensional Gaussian PDF
  - Gaussian mixture models (GMM)

PDF Modeling for Classification

- Procedure for classification based on PDF
  - Training stage: PDF modeling of each class based on the training dataset
  - Test stage: For each entry in the test dataset, pick the class with the max. PDF
- Commonly used classifiers:
  - Quadratic classifiers, with n-dim. Gaussian PDF
  - Gaussian-mixture-model classifier, with GMM PDF

1D Gaussian PDF Modeling

- 1D Gaussian PDF:

$$f(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- MLE of μ (mean) and σ^2 (variance)

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

D-dim. Gaussian PDF Modeling

- D-dim Gaussian PDF:

$$g(x;\mu,\Sigma)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

- MLE of μ (mean) and Σ (covariance matrix)

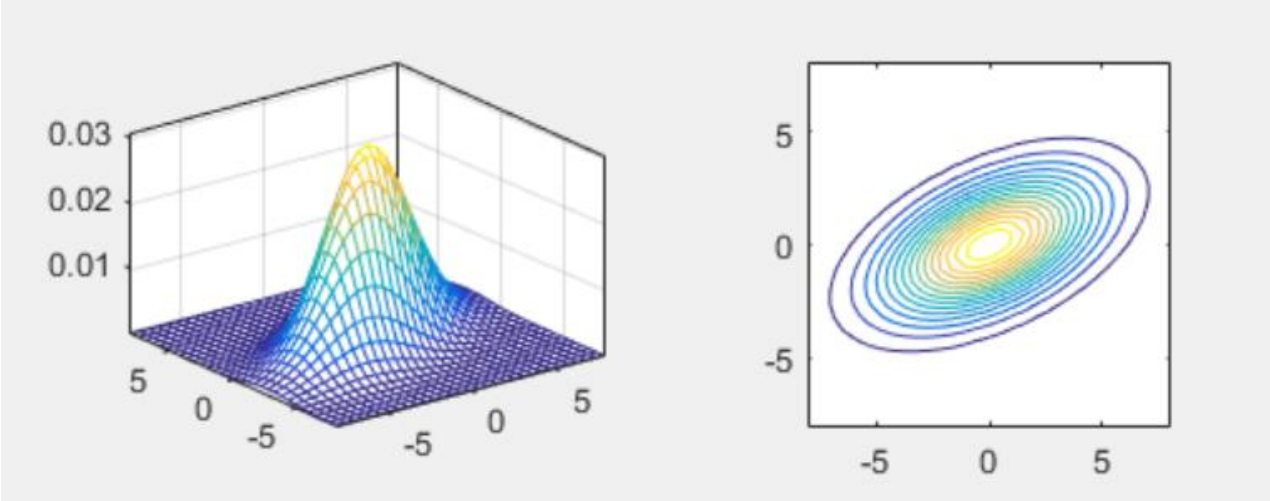
$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})(x_i-\hat{\mu})^T$$

2D Gaussian PDF

- Bivariate normal distribution:

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 9 & 3 \\ 3 & 4 \end{bmatrix}$$



PDF **Contours**

Training and Test Stages of QC

Quiz!

- Training stage
  - Identify the Gaussian PDF of each class via MLE
- Test stage
  - Assign a sample point to the class C by taking class prior into consideration:

$$\hat{C} = \arg\max_{C} \Pr(C) * pdf_{C}(\mathbf{x})$$

Prior!

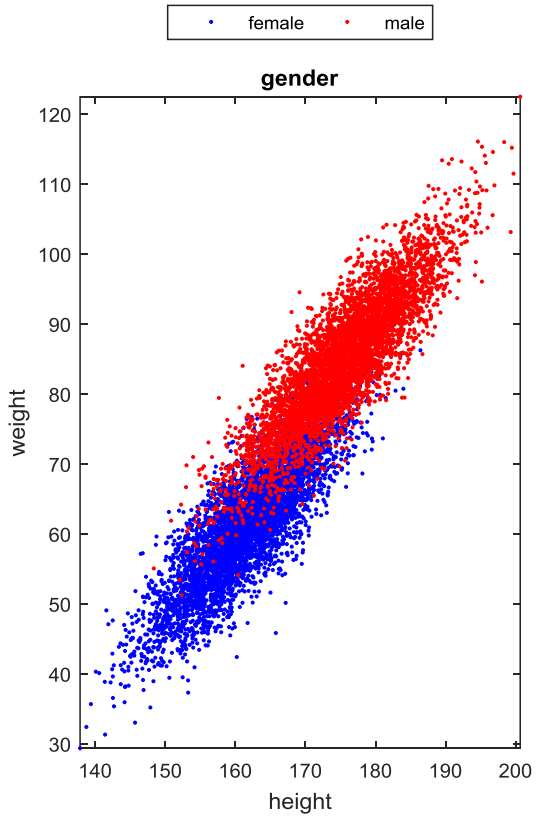
Characteristics of QC

- If each class is modeled by a Gaussian PDF, the decision boundary between any two classes is a quadratic curve.
  - That is why it is called quadratic classifier.
  - How to prove it?

Quiz!

QC on Gender Dataset (1/2)

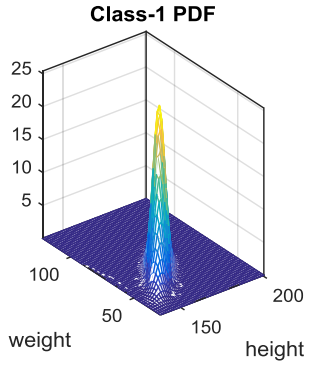
- Scatter plot of Gender dataset

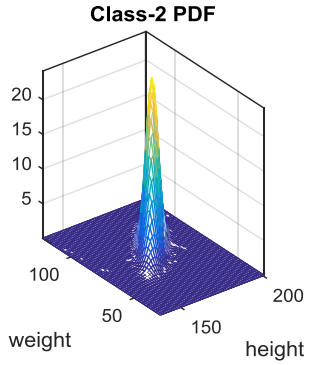


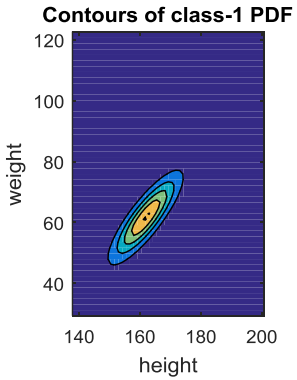

```
ds=prData('gender');
figure; dsScatterPlot(ds);
```

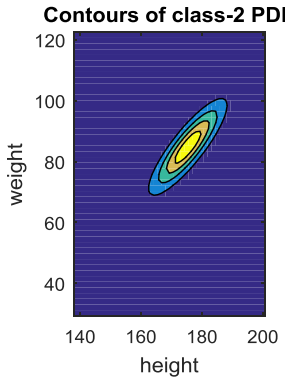
QC on Gender Dataset (2/2)

- PDF for each class



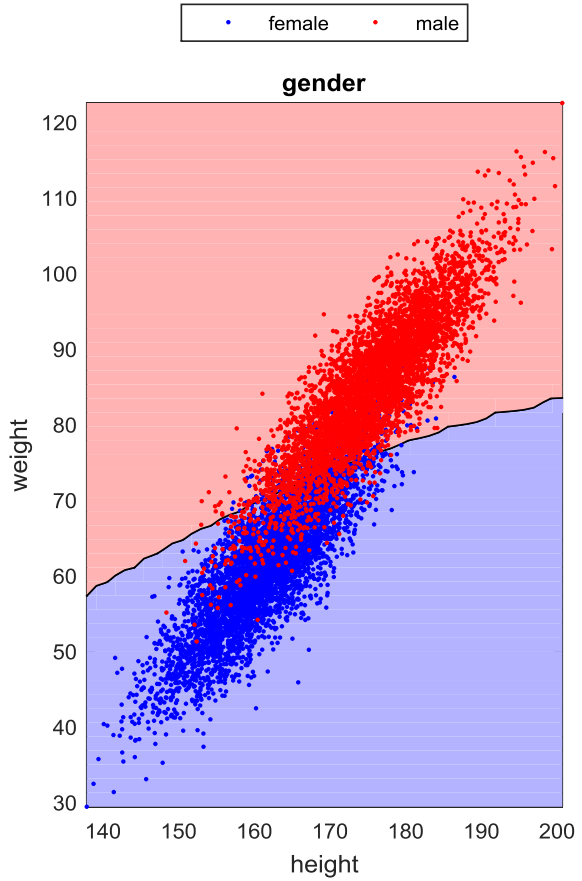






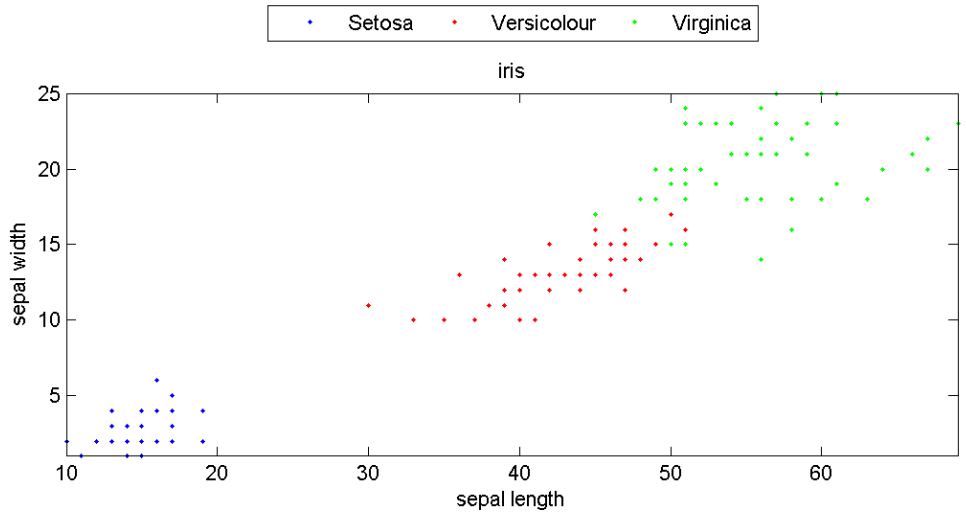

```
ds=prData('gender');
[qcPrm, logLike, recogRate, hitIndex]=qcTrain(ds);
figure; qcPlot(ds, qcPrm, '2dPdf');
figure; qcPlot(ds, qcPrm, 'decBoundary');
```

- Decision boundary



QC on Iris Dataset (1/2)

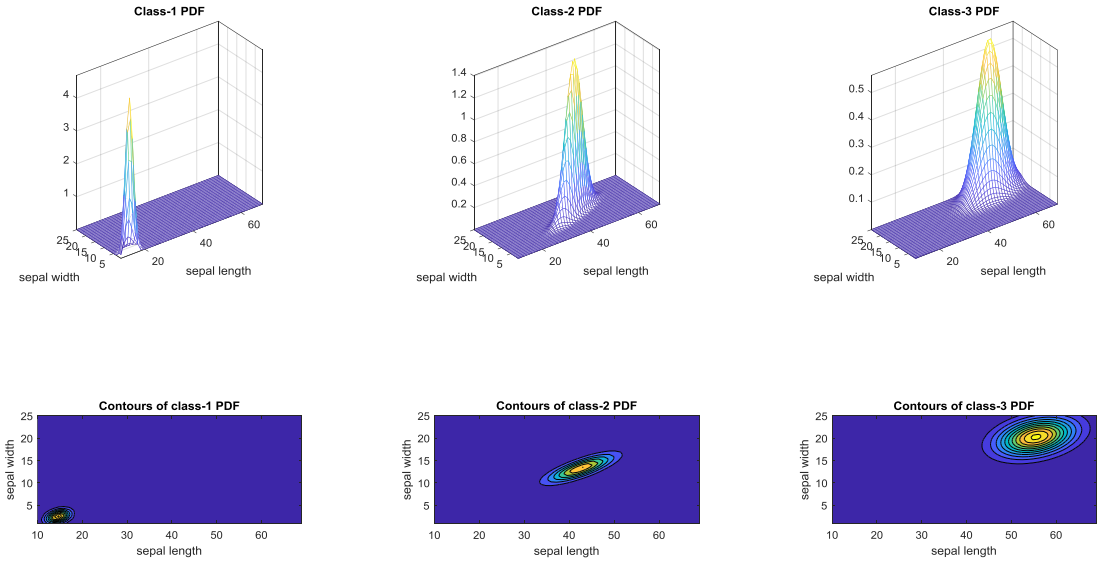
- Scatter plot of Iris dataset (with only the last two dim.)




```
ds=prData('iris');
ds.input=ds.input(3:4, :);
figure; dsScatterPlot(ds);
```

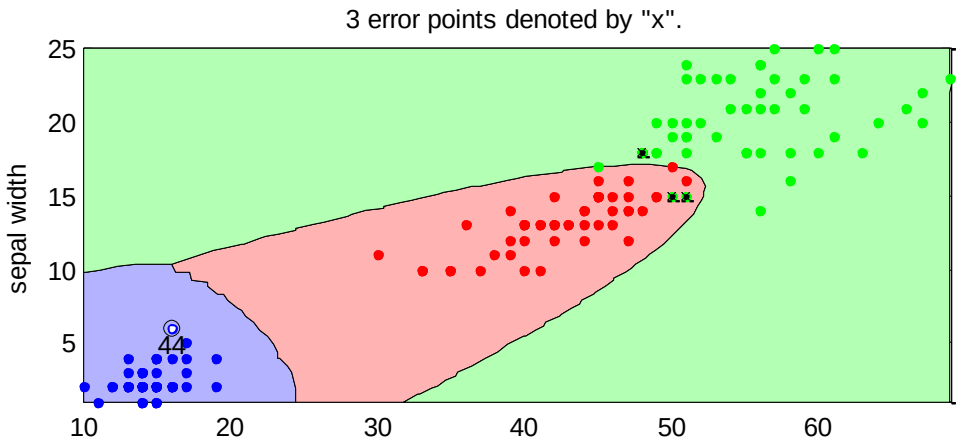
QC on Iris Dataset (2/2)

- PDF for each class



- Dec. boundaries

```
ds=prData('iris');
ds.input=ds.input(3:4, :);
[qcPrm, logLike, recogRate, hitIndex]=qcTrain(ds);
figure; qcPlot(ds, qcPrm, '2dPdf');
ds.hitIndex=hitIndex;        % For plotting
figure; qcPlot(ds, qcPrm, 'decBoundary');
```

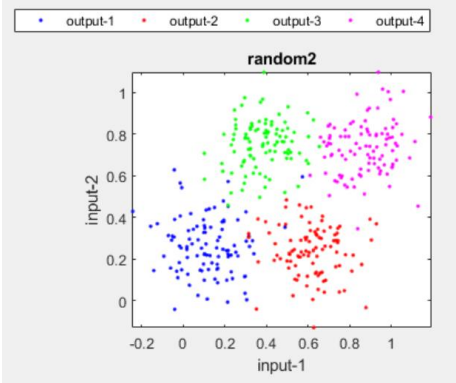


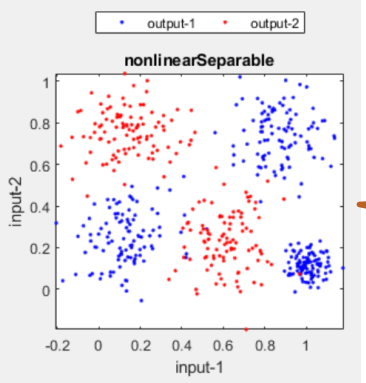
Strength and Weakness of QC

Quiz!

- Strength
  - Easy computation when the dimension d is small
  - Efficient way to compute leave-one-out cross validation
- Weakness
  - The covariance matrix (d by d) is big when the dimension d is median large
  - The inverse of the covariance matrix may not exist
  - Cannot handle bi-modal data correctly

Uni-modal!





Multi-modal!

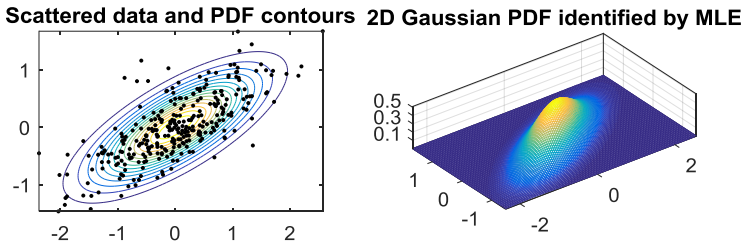
Modified Versions of QC

Quiz!

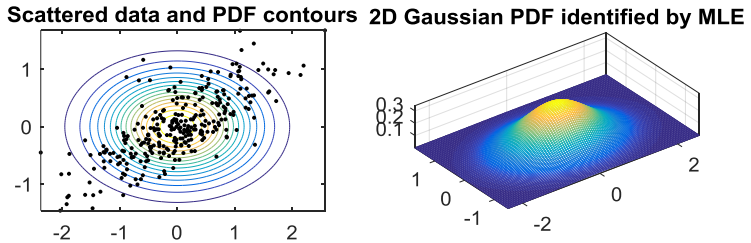
- How to modify QC such that it won't deal with a big covariance matrix?
  - Make the covariance matrix diagonal ➔ Equivalent to naïve Bayes classifiers (proof?)
  - Make the covariance matrix a constant times an identity matrix

gaussianMle;

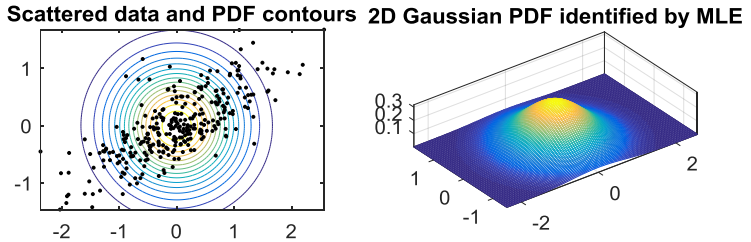
Σ: full



Σ: diagonal



$\Sigma = \sigma^2 I$



Naive Bayes Classifiers (NBC)

J.-S. Roger Jang (張智星)
jang@mirlab.org
http://mirlab.org/jang
MIR Lab, CSIE Dept.
National Taiwan University

Assumptions & Characteristics

- Assumptions:
  - Statistical independency between features
  - Statistical independency between samples
  - Each feature governed by a feature-wise parameterized PDF (usually a 1D Gaussian)
- Characteristics
  - Simple and easy (That's why it's named "naive".)
  - Highly successful in real-world applications regardless of the strong assumptions

Training and Test Stages of NBC

Quiz!

- Training stage
  - Identify class PDF, as follows.
    - Identify feature PDF by MLE for 1D Gaussians
    - Class PDF is the product of all the corresponding feature PDFs

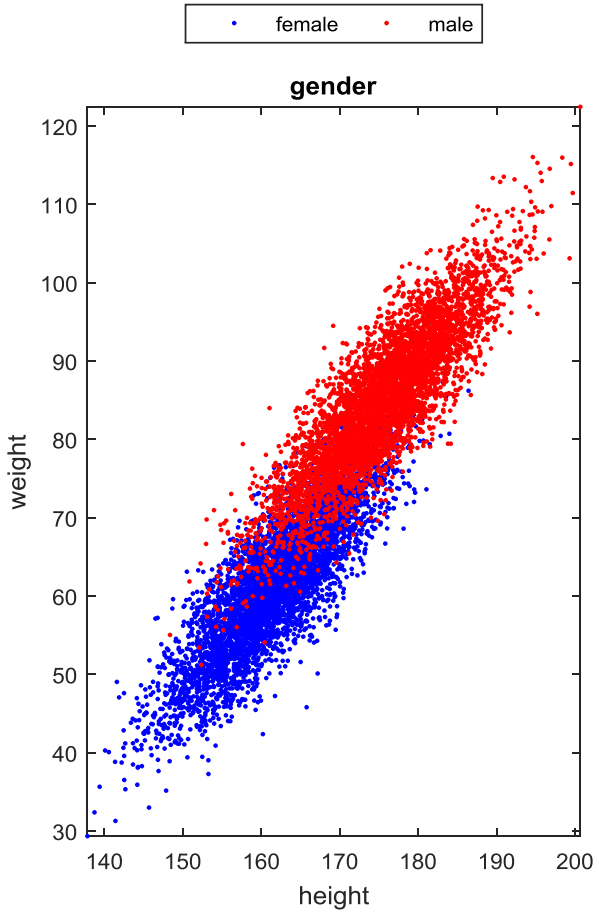
$$pdf_C(\mathbf{x}) = pdf_{1,C}(x_1) * pdf_{2,C}(x_2) * \cdots * pdf_{d,C}(x_d)$$

- Test stage
  - Assign a sample to the class by taking class prior into consideration:

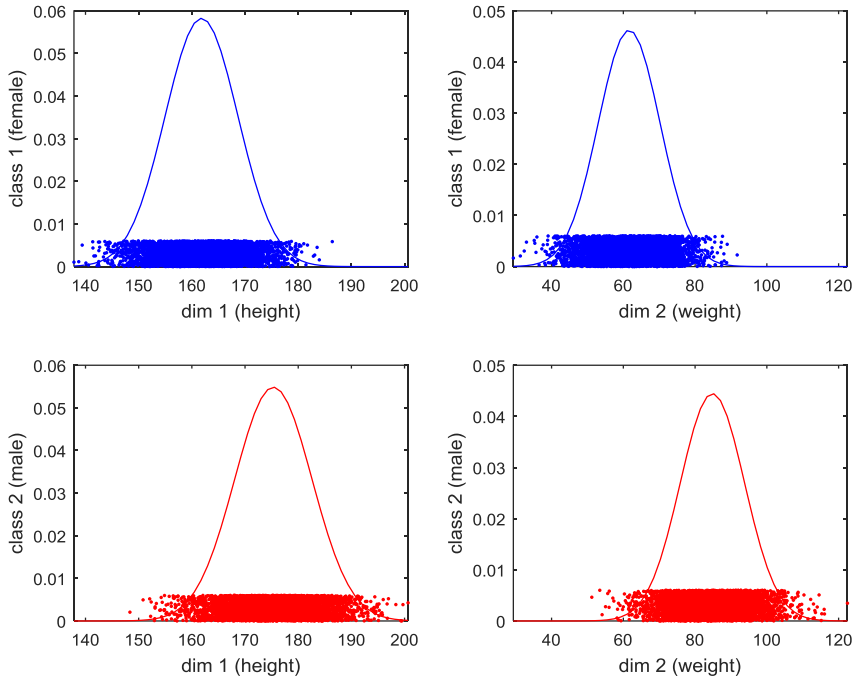
$$\hat{C} = \arg\max_C \Pr(C) * pdf_C(\mathbf{x})$$

NBC for Gender Dataset (1/2)

- Scatter plot of Gender dataset



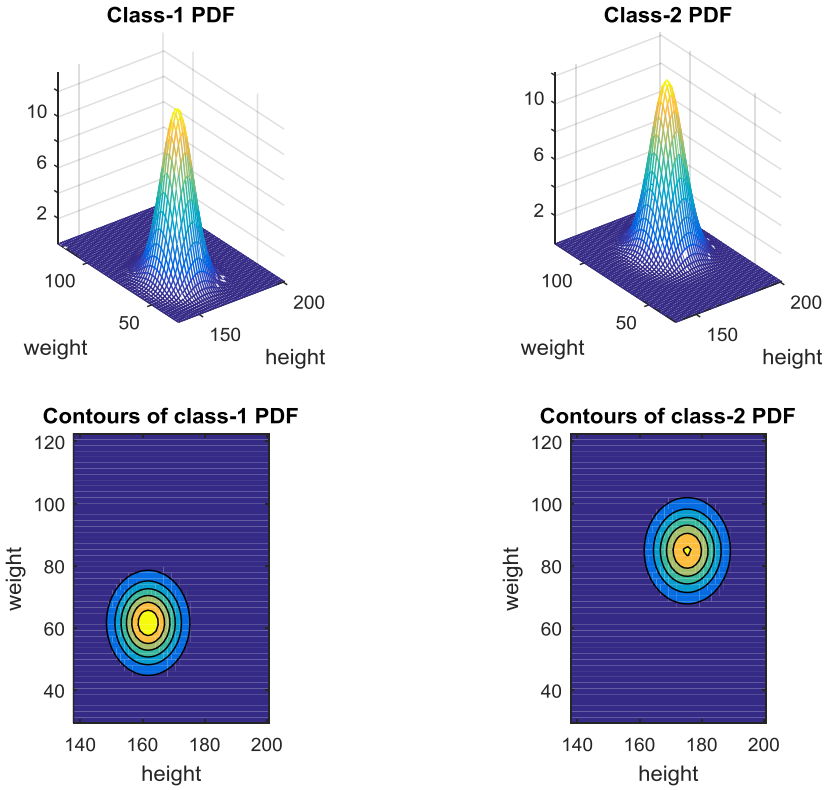
- PDF on each features and each class




```
ds=prData('gender');
figure; dsScatterPlot(ds);
[nbcPrm, logLike, recogRate, hitIndex]=nbcTrain(ds);
figure; nbcPlot(ds, nbcPrm, '1dPdf');
```

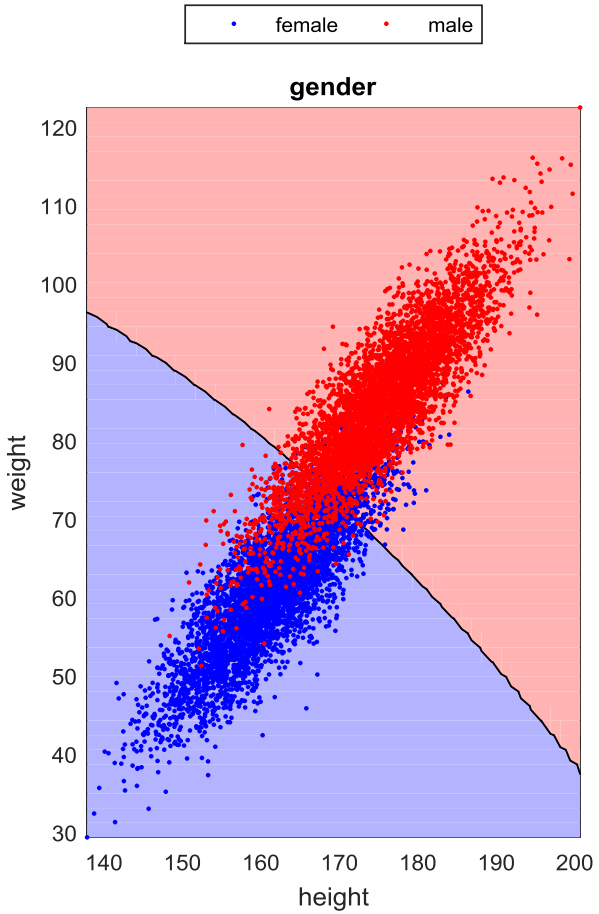
NBC for Gender Dataset (2/2)

- PDF for each class



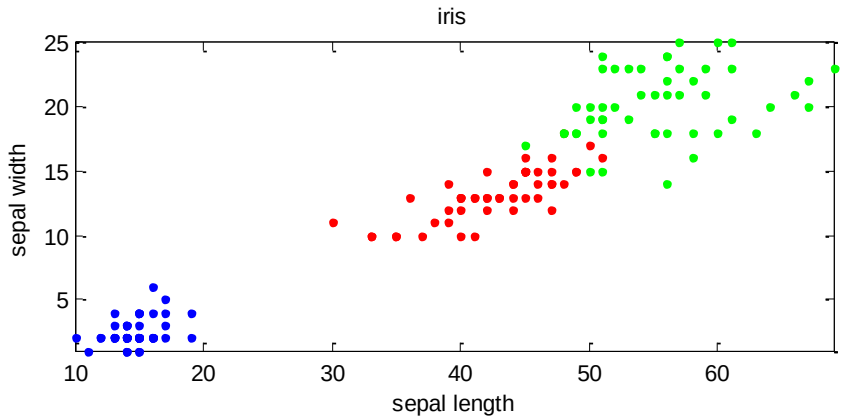

```
ds=prData('gender');
[nbcPrm, logLike, recogRate, hitIndex]=nbcTrain(ds);
figure; nbcPlot(ds, nbcPrm, '2dPdf');
figure; nbcPlot(ds, nbcPrm, 'decBoundary');
```

- Decision boundary



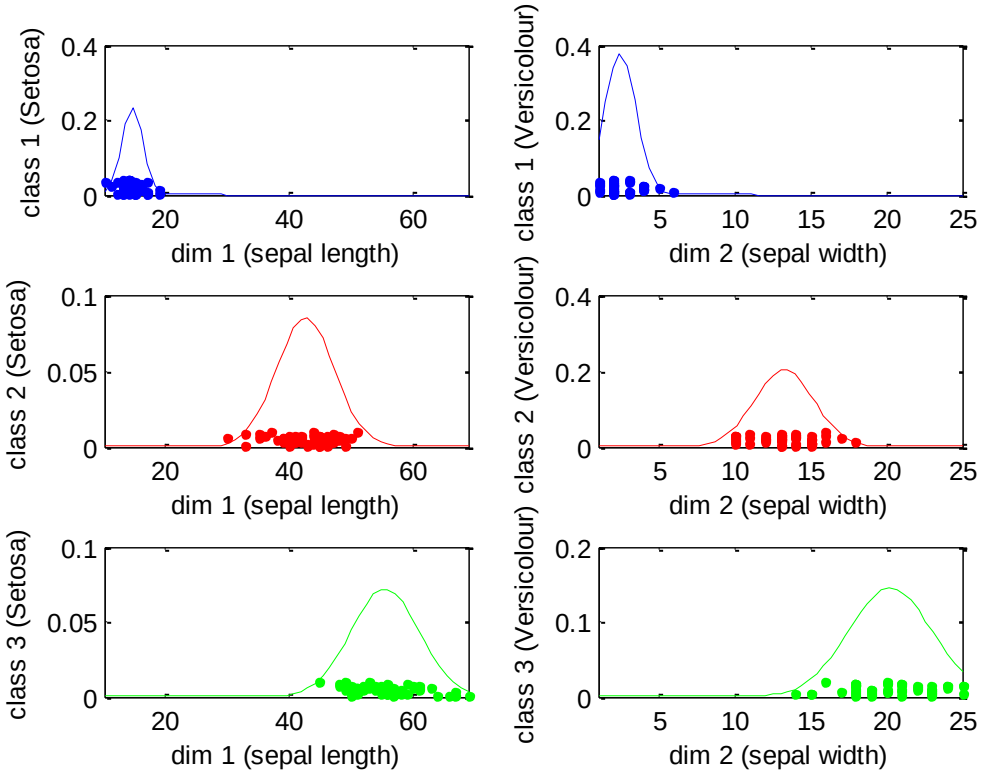
NBC for Iris Dataset (1/2)

- Scatter plot of Iris dataset (with only the last two dim.)



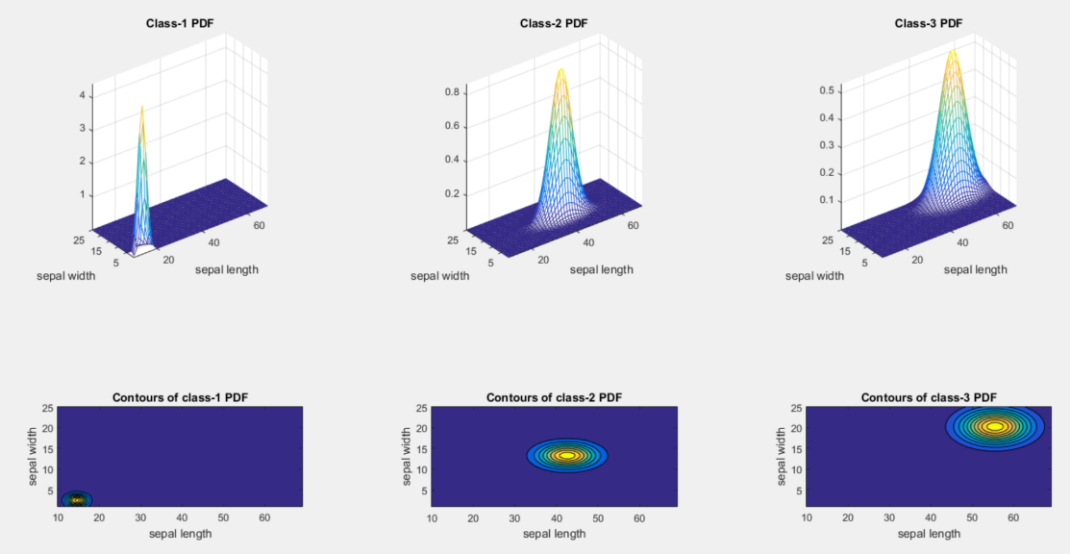

```
ds=prData('iris');
ds.input=ds.input(3:4, :);
figure; dsScatterPlot(ds);
[nbcPrm, logLike, recogRate, hitIndex]=nbcTrain(ds);
figure; nbcPlot(ds, nbcPrm, '1dPdf');
```

- PDF on each features and each class



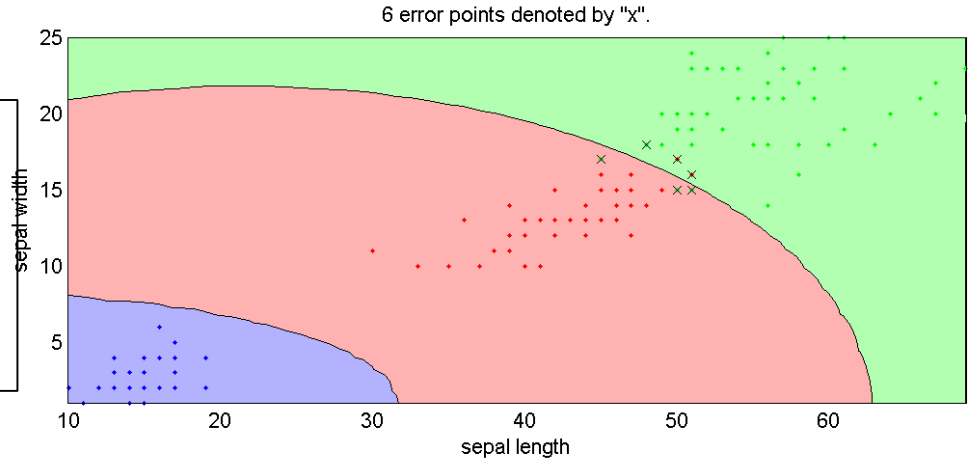
NBC for Iris Dataset (2/2)

- PDF for each class



- Dec. boundaries

```
ds=prData('iris');
ds.input=ds.input(3:4, :);
[nbcPrm, logLike, recogRate, hitIndex]=nbcTrain(ds);
figure; nbcPlot(ds, nbcPrm, '2dPdf');
ds.hitIndex=hitIndex;       % For plotting
figure; nbcPlot(ds, nbcPrm, 'decBoundary');
```



Strength and Weakness of NBC

Quiz!

- Strength
  - Fast computation during training and evaluation
  - Robust than QC
- Weakness
  - No able to deal with bi-modal data correctly
  - Class boundary not as complex as QC