機械学習特論

~理論とアルゴリズム~

第5回

(Linear Discriminant Analysis and classification measures)

講師 : 西郷浩人

Ex. 1

- Automatically discriminate the numbers by the 7-th subject. Which digit is misclassified to which one?
  - T(:,7,:)
- Complete the following table, where the largest number corresponds to the prediction by LDA.

prediction \ True	0	1	2	3	4	5	6	7	8	9	0
1		7296									
2		7268									
3		7267									
4		7270									
5		7261									
6		7248									
7		7271									
8		7270									
9		7278									
0		7244									

Hint

-

```
%Computation of a common covariance matrix
[a b c] = size(X);
S = zeros(a);
mu = zeros(a,c);
for i=1:c
 mu(:,i) = mean(X(:,:,i),2);
 S = S+cov(X(:,:,i)');
end
S = S ./ c;
invS = inv(S);
```

Answer

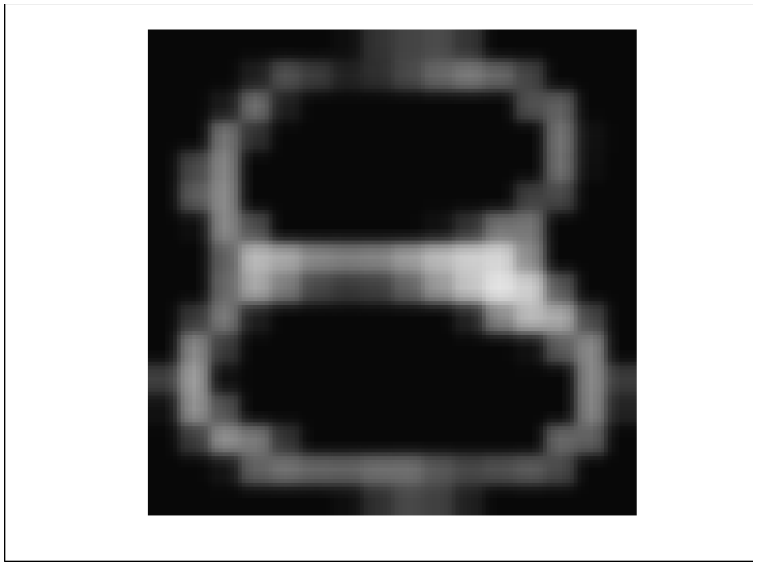
```
for i=1:10
 for j=1:10
  p(j,i)=T(:,7,i)'*invS*mu(:,j)-mu(:,j)'*invS*mu(:,j)./2;
 end
end
[M I] = max(p)

%conversion to probabilities
p=p./sum(p)
```

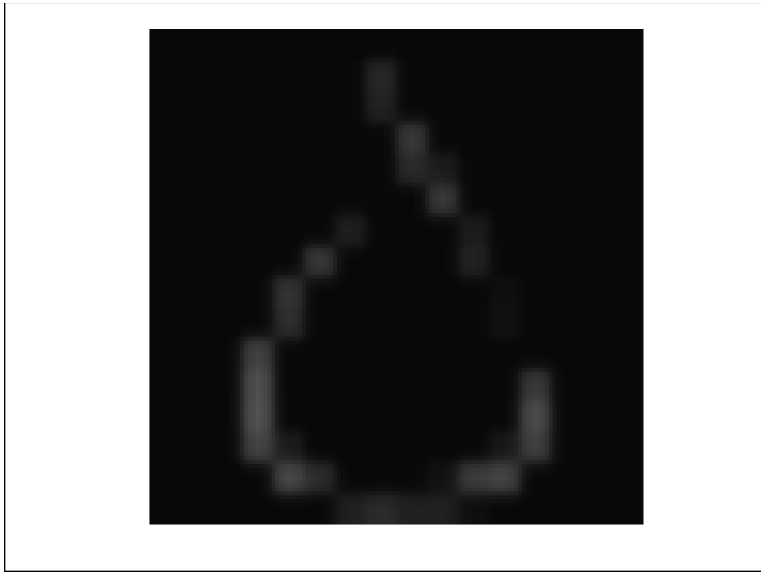
Answer

P =

prediction \ True									
0.100396	0.099725	0.099608	0.099904	0.099744	0.099741	0.099728	0.099602	0.099686	0.099848
0.100013	0.100148	0.099924	0.099943	0.099995	0.100116	0.100077	0.100036	0.099951	0.100070
0.100001	0.100026	0.100291	0.099965	0.099952	0.099957	0.100076	0.100254	0.100064	0.100182
0.100034	0.100114	0.100052	0.100267	0.099997	0.099973	0.100055	0.100041	0.100154	0.099989
0.099908	0.100028	0.100175	0.099925	0.100253	0.100099	0.099823	0.100096	0.099951	0.100032
0.099732	0.100058	0.100033	0.099808	0.100104	0.100420	0.099864	0.099936	0.099767	0.100075
0.100053	0.099922	0.099816	0.100116	0.099803	0.099785	0.100392	0.099917	0.100189	0.099854
0.100036	0.100065	0.100098	0.100110	0.100211	0.099957	0.099973	0.100084	0.100072	0.099998
0.100144	0.100020	0.099914	0.100235	0.099933	0.099846	0.100212	0.100045	0.100337	0.099927
0.099685	0.099893	0.100090	0.099728	0.100010	0.100106	0.099800	0.099988	0.099828	0.100025



Misclassify 8 as 3



Misclassify 0 as 3

Properties of LDA

Example in binary classification

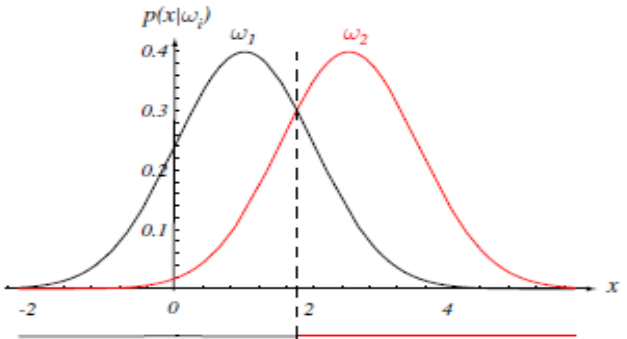
$$f(\boldsymbol{x}) = \boldsymbol{x}'\Sigma^{-1}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=2}) - \frac{1}{2}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=2})'\Sigma^{-1}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=2}) + \log\frac{n_{y=1}}{n_{y=2}}$$

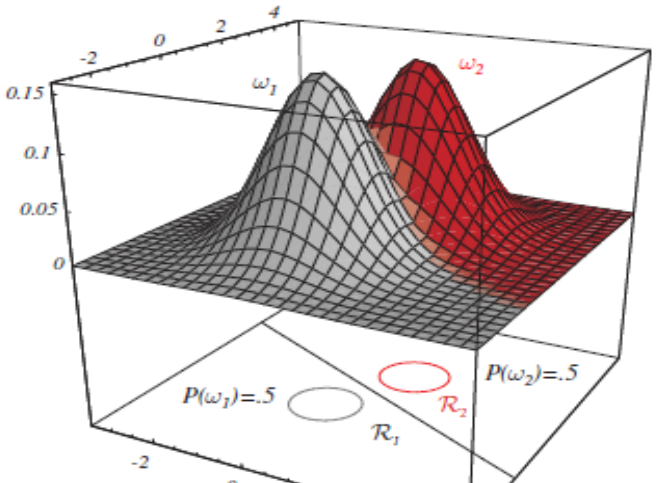
- This discriminant function can be written by setting **a** and b as

$$\boldsymbol{a} = \Sigma^{-1}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=2})$$

$$b = \frac{-1}{2}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=2})'\Sigma^{-1}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=2}) + \log\frac{n_{y=1}}{n_{y=2}}$$

then $f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{a} + b$ is linear with respect to x. The corresponding separating hyperplane is linear and orthogonal to the discriminant function.



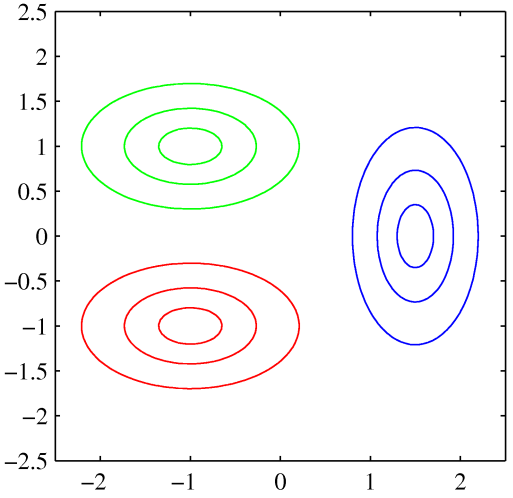


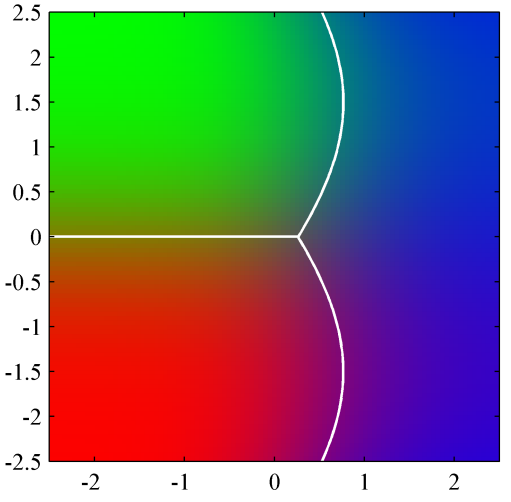
Properties of LDA

- Discriminant function is orthogonal to the line passing through the means of the two classes.

$$f(\boldsymbol{x}) = \boldsymbol{x}' \Sigma^{-1}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=2}) - \frac{1}{2}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=2})' \Sigma^{-1}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=2}) + \log \frac{n_{y=1}}{n_{y=2}}$$

- In the figure, red and green class has the common covariance matrix, and their separating hyperplane is *linear*.
- Blue class has a different covariance matrix, and its separating hyperplane is *nonlinear.*





When covariance matrix is not common

$$\log p(y|\boldsymbol{x}) = \frac{1}{2}\log(|\Sigma_y|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_y)'\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_y) + \log(n_y) + C$$

- Consider taking difference between the two classes.

$$\log p(y=1|\boldsymbol{x}) = \frac{1}{2}\log(|\Sigma_{y=1}|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{y=1})'\boldsymbol{\Sigma}_{y=1}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{y=1}) + \log(n_{y=1}) + C$$

$$\log p(y=-1|\boldsymbol{x}) = \frac{1}{2}\log(|\Sigma_{y=-1}|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{y=-1})'\boldsymbol{\Sigma}_{y=-1}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{y=-1}) + \log(n_{y=-1}) + C$$

$$f(\boldsymbol{x}) = \log p(y=1|\boldsymbol{x}) - \log p(y=-1|\boldsymbol{x})$$

$$= \frac{-1}{2}\left(\boldsymbol{x}'(\Sigma_{y=1}^{-1} - \Sigma_{y=-1}^{-1})\boldsymbol{x} + 2\,\boldsymbol{x}'(\Sigma_{y=1}^{-1}\boldsymbol{\mu}_{y=1} - \Sigma_{y=-1}^{-1}\boldsymbol{\mu}_{y=-1})\right) + C$$

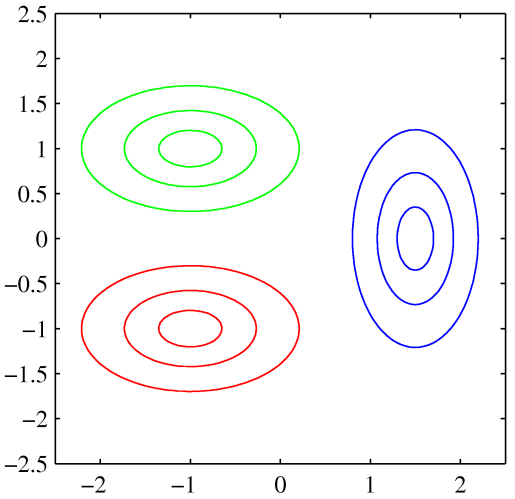
- where

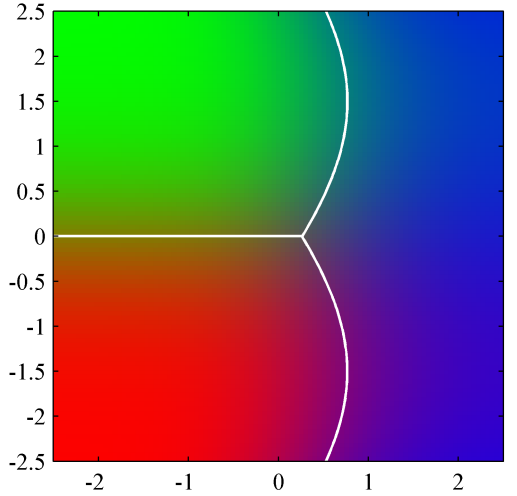
$$C = \frac{1}{2}\log\frac{|\Sigma_{y=1}|}{|\Sigma_{y=-1}|} + \log\frac{n_{y=1}}{n_{y=-1}} - \frac{1}{2}\left(\boldsymbol{\mu}_{y=1}'(\Sigma_{y=1}^{-1} - \Sigma_{y=-1}^{-1})\boldsymbol{\mu}_{y=-1}\right)$$

LDA and QDA

$$f(\boldsymbol{x}) = \frac{-1}{2}\left(\boldsymbol{x}'\left(\Sigma_{y=1}^{-1} - \Sigma_{y=-1}^{-1}\right)\boldsymbol{x} + 2\,\boldsymbol{x}'\left(\Sigma_{y=1}^{-1}\boldsymbol{\mu}_{\boldsymbol{y=1}} - \Sigma_{y=-1}^{-1}\boldsymbol{\mu}_{\boldsymbol{y=-1}}\right)\right) + C$$

- In the above discriminant function, when covariance matrix is not common, the first term does not vanish, and results in a function which is quadratic with respect to **x**.
  - Between red and blue, or green and blue in the figure below.
- In this case, the resulting classification rule is called as Quadratic Discriminant Analysis (QDA).

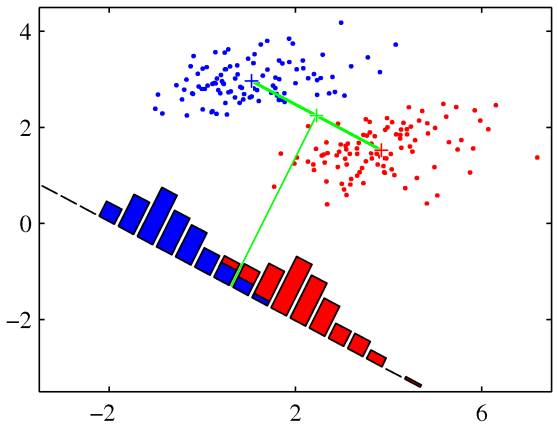




Yet another derivation of LDA (Fishers' approach)

$$f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{a} + b$$

- We have learned that in binary classification case, discriminant function becomes linear if the common covariance assumption holds.
- Here we consider the opposite; assuming the linearity of the discriminant function, then derive LDA by maximizing the distance between the class centers.
  - We determine the weight **a** first, then bias term b later.



Binary classification case

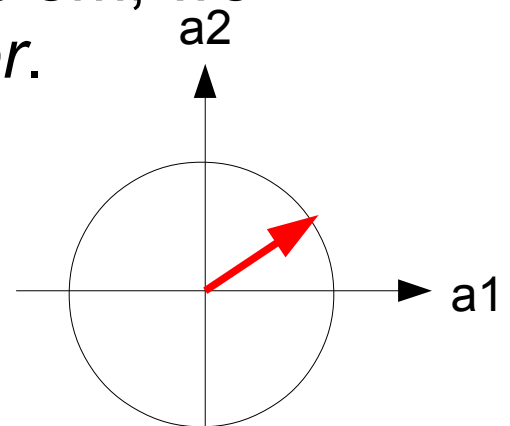
- First, set the class centers as below.

$$\boldsymbol{\mu}_1 = \frac{1}{N_1}\sum_{n\in C_1} \boldsymbol{x}_n \qquad \boldsymbol{\mu}_2 = \frac{1}{N_2}\sum_{n\in C_2} \boldsymbol{x}_n$$

- Then consider finding the weight **a** that maximizes the distance between the class centers.

$$max_{\boldsymbol{a}}\, \boldsymbol{a}'(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$
$$s.t.\ \boldsymbol{a}'\boldsymbol{a} = 1$$

- Note that the objective function can be maximized arbitrary large without constraints on **a**, so we limit its length to 1.
  - In the two dimensional case, **a** is located on a circle.
- In order to solve constrained optimization problem, we make use of the *method of Lagrange multiplier*.



Method of Lagrange multiplier (ラグランジュ乗数法)

$$max_{\boldsymbol{x}} f(\boldsymbol{x}) \quad s.t.\, g(\boldsymbol{x}) = 0$$

- A method for solving constrained optimization problem.
  - Newton method or gradient descent cannot be applied directly because of constraints.
- So we consider solving a new objective function L that additionally includes constraint.

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda g(\boldsymbol{x})$$

  - λ is called a Lagrangian multiplier.
- L is a function of **x** and λ, so we try to find a local maximum by taking derivatives of L with respect to L and λ, and setting them to 0s.

$$\frac{\partial L(\boldsymbol{x}, \lambda)}{\partial \lambda} = 0 \qquad \frac{\partial L(\boldsymbol{x}, \lambda)}{\partial \boldsymbol{x}} = \mathbf{0}$$

example

- Consider the following constrained optimization problem.

$$max_x f(\boldsymbol{x}) \quad s.t.\, g(\boldsymbol{x}) = 0$$

$$f(\boldsymbol{x}) = 1 - x_1^2 - x_2^2 \qquad g(\boldsymbol{x}) = x_1 + x_2 - 1 = 0$$

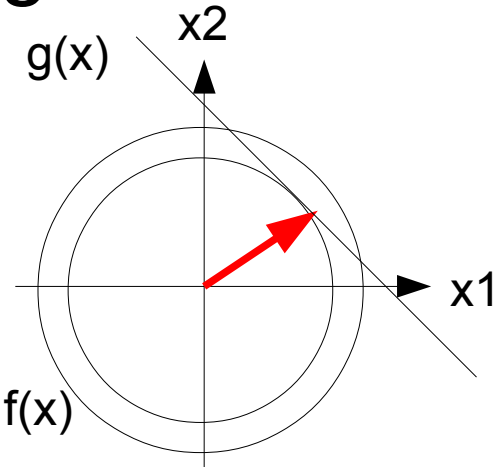
- Let the new objective function be L, and set it to zero.

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda\, g(\boldsymbol{x}) = 1 - x_1^2 - x_2^2 - \lambda(x_1 + x_2 - 1)$$

$$\frac{\partial L(\boldsymbol{x}, \lambda)}{\partial \lambda} = x_1 + x_2 - 1 = (0) \qquad \frac{\partial L(x_1, \lambda)}{\partial x_1} = -2x_1 - \lambda = (0) \qquad \frac{\partial L(x_2, \lambda)}{\partial x_2} = -2x_2 - \lambda = (0)$$

- By solving a system of equations, we get the following solution.

$$\lambda = -1, \quad x_1 = \frac{1}{2}, \quad x_2 = \frac{1}{2}$$



Solving binary classification problem by LDA

$$max_{\boldsymbol{a}}\, \boldsymbol{a}'(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$
$$s.t.\ \boldsymbol{a}'\boldsymbol{a}=1$$

- Using a method of Lagrange multiplier, we get the following equivalent problem.

$$max\ L=max_{\boldsymbol{a},\lambda}\, \boldsymbol{a}'(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)-\lambda(\boldsymbol{a}'\boldsymbol{a}-1)$$

- Taking derivative of L with respect to a and λ, and setting them to zeros.

$$\frac{\partial L}{\partial \boldsymbol{a}}=\boldsymbol{\mu}_1-\boldsymbol{\mu}_2-2\lambda\boldsymbol{a}=0,\quad \frac{\partial L}{\partial \lambda}=\boldsymbol{a}'\boldsymbol{a}-1=0$$

$$\Leftrightarrow\quad \boldsymbol{a}=\frac{1}{2\lambda}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\quad \lambda=\frac{\sqrt{((\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2))}}{2}$$

$$\Leftrightarrow\quad \boldsymbol{a}=\frac{\boldsymbol{\mu}_1-\boldsymbol{\mu}_2}{\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|}$$

Vector Norms

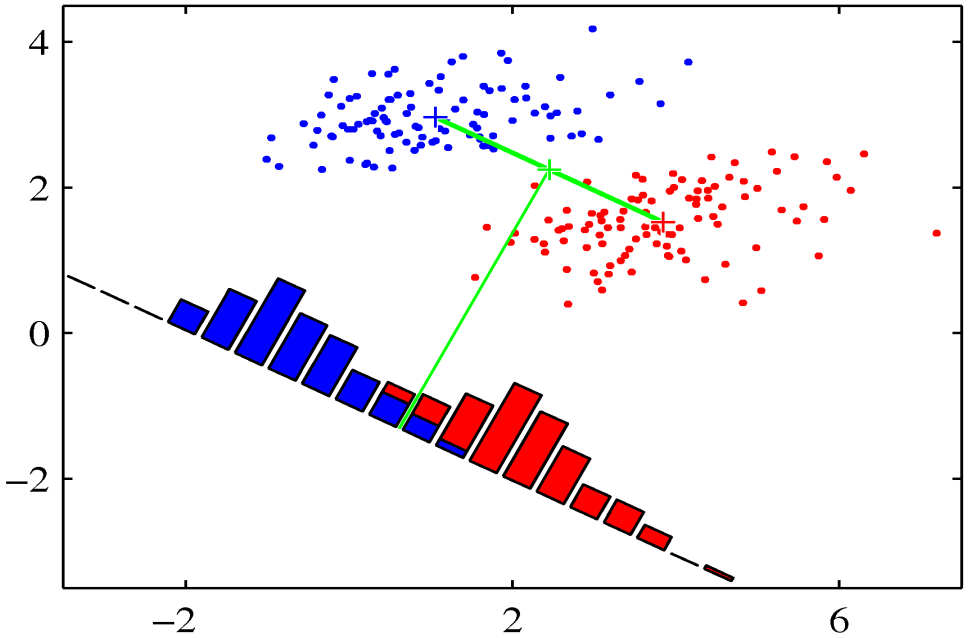
- definition $$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

  - p =1: 1-norm (Manhattan distance)
    $$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$
  - p = 2: 2-norm (Euclidian distance)
    $$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$
  - p = ∞ : infinity-norm (Chebyshev distance)
    $$\|x\|_\infty = max_i |x_i|$$

Problem in maximizing the distance between the class centers.

$$a = \frac{\mu_1 - \mu_2}{\|\mu_1 - \mu_2\|}$$

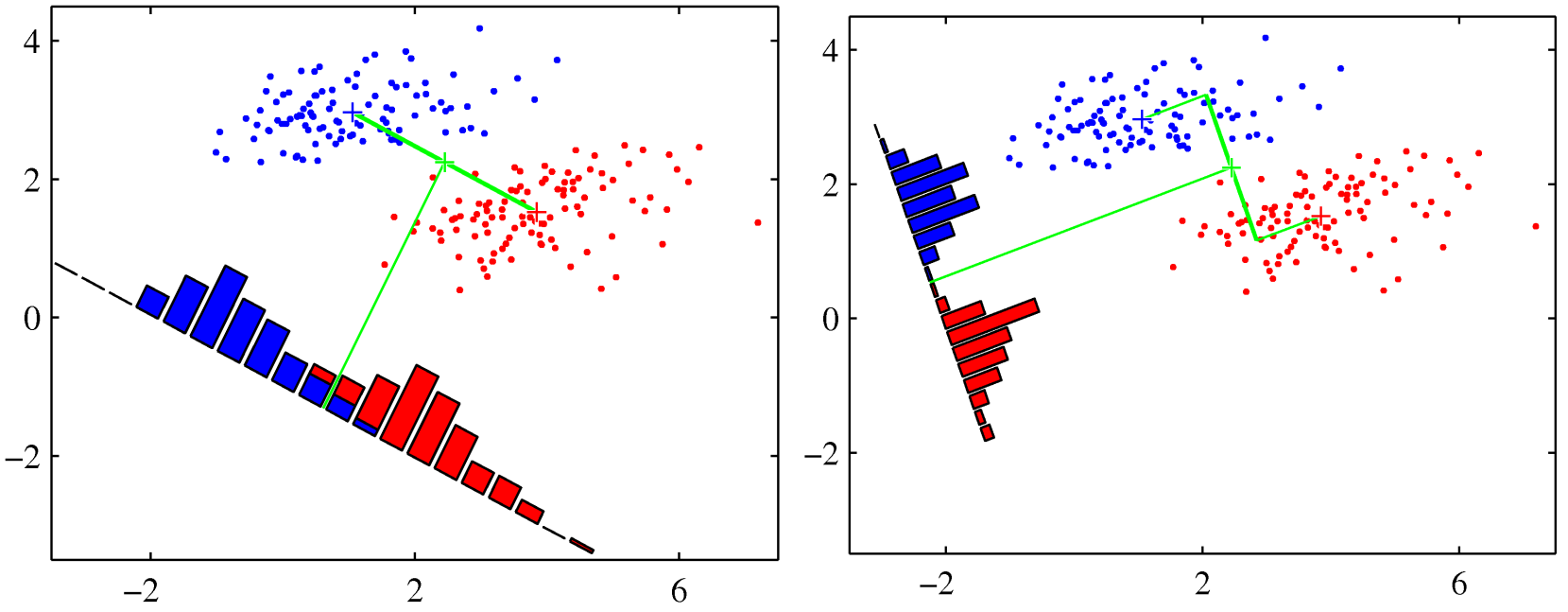
- Does not work on the following example



- The problem is that the superposition among the classes is too large when projected on a one-dimensional space, which is due to the large correlation among features in two classes.

Towards better Classification

- Not only maximizing the distance between the class centers, consider minimizing the variances in each class.



Fisher's criterion

- Let within-class variance be. (assuming $y_n = \boldsymbol{a}' \boldsymbol{x}_n$)

$$s_1^2 = \sum_{n \in C_1} (y_n - \boldsymbol{a}' \boldsymbol{\mu_1})^2 \quad s_2^2 = \sum_{n \in C_2} (y_n - \boldsymbol{a}' \boldsymbol{\mu_2})^2$$

- Then the sum of within class variance is obtained as $s_1^2 + s_2^2$

  and the between class variance (or the distance between the two centers) is given as $\|\boldsymbol{a}'(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})\|^2 = \boldsymbol{a}'(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{a}$

- The solution is given by maximizing the between class variance and minimizing the within class variance. So we consider maximizing the rate between them.

$$J(\boldsymbol{a}) = \frac{\|\boldsymbol{a}'(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})\|^2}{s_1^2 + s_2^2}$$

$$= \frac{\boldsymbol{a}'(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{a}}{\boldsymbol{a}'(\sum_{n \in C_1} (\boldsymbol{x}_n - \boldsymbol{\mu_1})(\boldsymbol{x}_n - \boldsymbol{\mu_1})' + \sum_{n \in C_2} (\boldsymbol{x}_n - \boldsymbol{\mu_2})(\boldsymbol{x}_n - \boldsymbol{\mu_2})')\boldsymbol{a}} = \frac{\boldsymbol{a}' \Sigma_B \boldsymbol{a}}{\boldsymbol{a}' \Sigma \boldsymbol{a}}$$

  - J is known as Fishers's criterion, or Rayleigh quotient, and often used for clustering or ANOVA.

Ex. 1

- By taking a derivative of $J(\boldsymbol{a}) = \frac{\boldsymbol{a}' \Sigma_B \boldsymbol{a}}{\boldsymbol{a}' \Sigma \boldsymbol{a}}$ with respect to a and setting it to zero, prove that the following equation holds.

$$(\boldsymbol{a}' \Sigma_B \boldsymbol{a}) \Sigma \boldsymbol{a} = (\boldsymbol{a}' \Sigma \boldsymbol{a}) \Sigma_B \boldsymbol{a}$$

  - you can use the following rule for a matrix/vector derivative.

$$\frac{\partial}{\partial \boldsymbol{a}} \boldsymbol{a}' \boldsymbol{X} \boldsymbol{a} = 2 \boldsymbol{X} \boldsymbol{a}$$

Deriving the weight a

- From Ex. 1, a vector a that maximizes J satisfies

$$(\boldsymbol{a}'\Sigma_B\boldsymbol{a})\Sigma\boldsymbol{a}=(\boldsymbol{a}'\Sigma\boldsymbol{a})\Sigma_B\boldsymbol{a}$$

- Since we are interested only in the direction of a, so scalars $(\boldsymbol{a}'\Sigma_B\boldsymbol{a})$ and $(\boldsymbol{a}'\Sigma\boldsymbol{a})$ can be ignored.

$$\Sigma\boldsymbol{a}\propto\Sigma_B\boldsymbol{a}$$

- Then we obtain the following relationship.

$$\boldsymbol{a}\propto\Sigma^{-1}(\mu_2-\mu_1)$$

Discriminant direction

$$\boldsymbol{a} \propto \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

- Coincides with the derivation of LDA in the previous lecture.

$$f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{a} + b$$

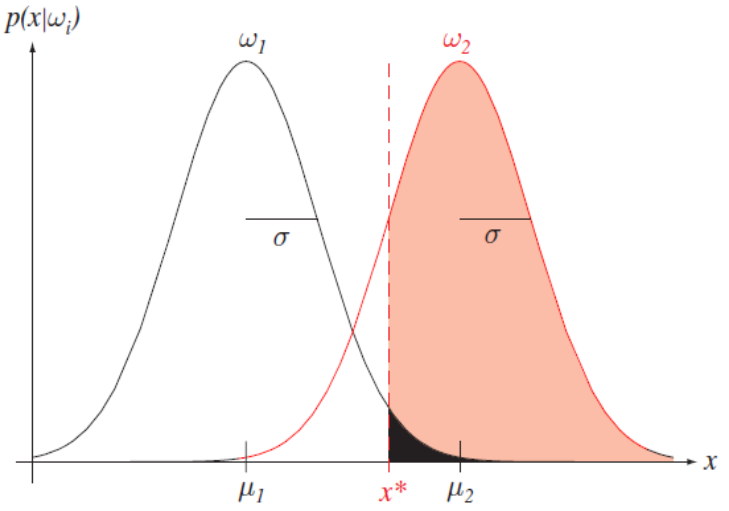
$$\boldsymbol{a} = \Sigma^{-1}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=2})$$

$$b = \frac{-1}{2}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=2})'\Sigma^{-1}(\boldsymbol{\mu}_{y=1} - \boldsymbol{\mu}_{y=2}) + \log\frac{n_{y=1}}{n_{y=2}}$$

Classification measures (binary case)

2 types of truth and errors

- Predicting the possession of a disease by genetic diagnosis.
- In the figure below, if we use separating hyperplane marked by x*, then we get the two types of truth;
  - Predicting a sick person as sick.
  - Predicting a healthy person as healthy.
- In the figure below, if we use separating hyperplane marked by x*, then we get the two types of truth;
  - Predicting a sick person as healthy.
  - Predicting a healthy person as sick.



example

The following table is often called as contingency table.

		actual: Yes	actual: No
prediction	Yes	80 **(True Positive)**	20 **(False Positive)**
	No	10 **(False Negative)**	90 **(True Negative)**

Evaluating classifier performance based on Contingency Table

.

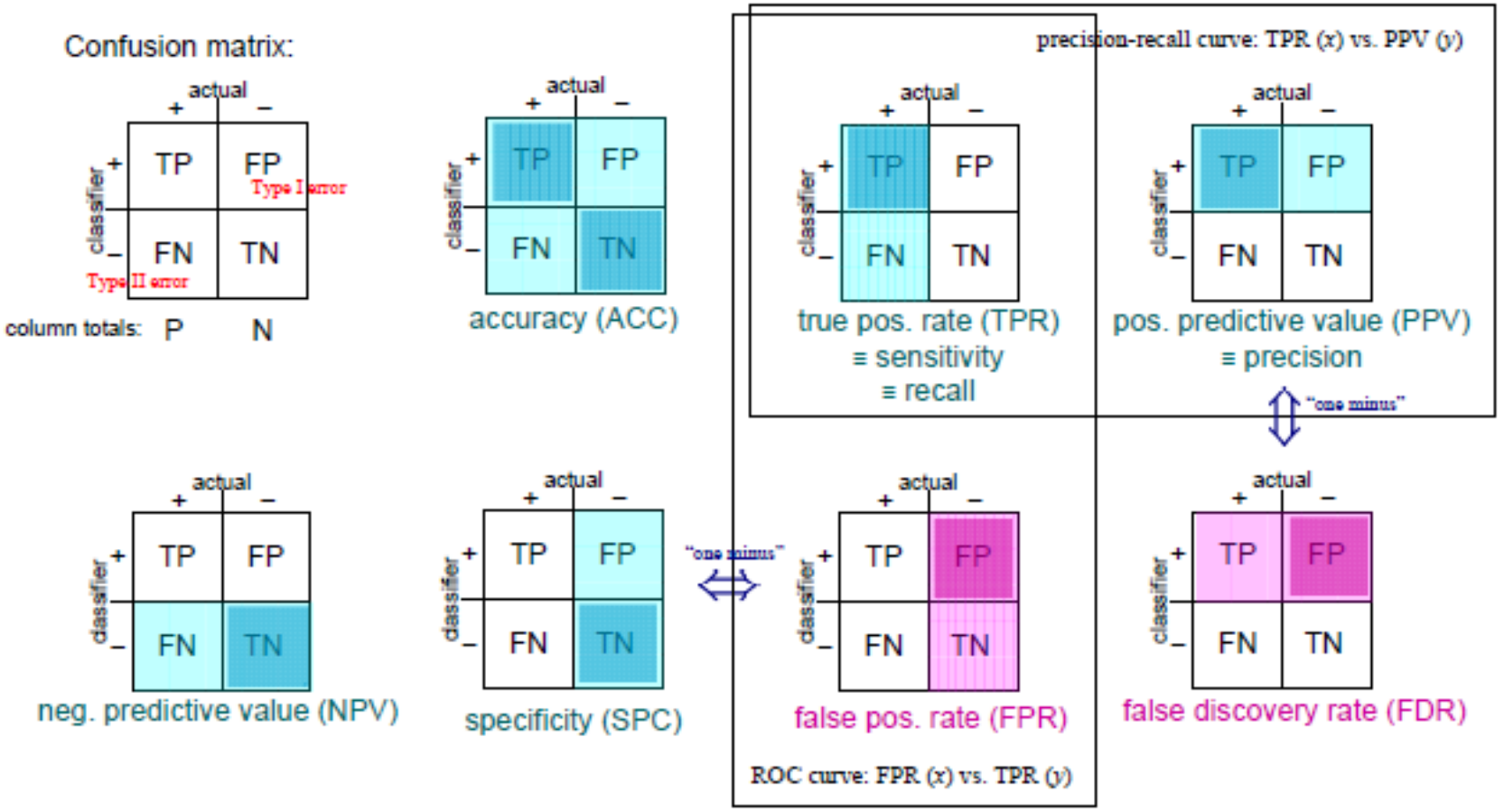
actual

prediction	Yes(Class1)	No(Class2)
Yes(Class1)	80 **(True Positive)**	20 **(False Positive)**
No(Class2)	10 **(False Negative)**	90 **(True Negative)**

Accuracy: ACC = (TP+TN)/(TP+FP+TN+FN) = (TP+TN) / All = 170/200
False Positive Rate : FPR = FP/(FP+TN) = 20/110
True Positive Rate(a.k.a. sensitivity, recall): TPR = TP/(TP+FN)=80/90
Positive Predictive Value(a.k.a. precision): PPV = TP/(TP+FP)=80/100
Specificity: TN/(FP+TN) = 90/110

Other measures

"Cheat sheet" on accuracy, precision, recall, TPR, FPR, specificity, sensitivity, ROC, and all that stuff!
William H. Press, ver 1.0, 3/29/08



value (between 0 and 1) = numerator / denominator
numerator = dark color shade
denominator = dark + light color shade
blue: value 1 is good
pink: value 0 is good

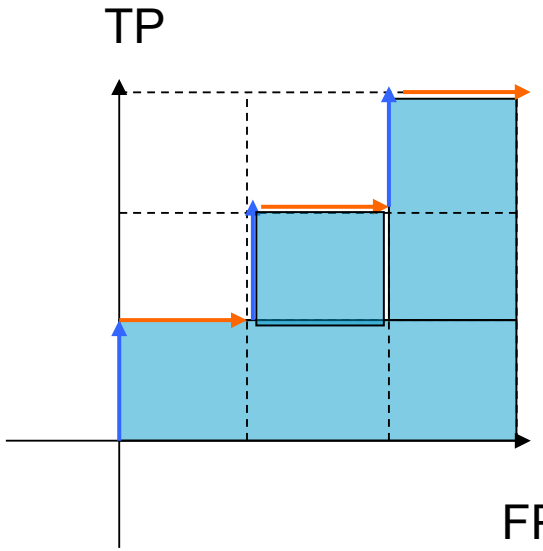
Map points from ROC to Precision-Recall or vice-versa: (TPR same values in both)

$$PPV = \frac{P\ TPR}{P\ TPR + N\ FPR} \qquad \textit{(ROC to P-R)}$$

$$FPR = \frac{P\ (1 - PPV)\ TPR}{N\ PPV} \qquad \textit{(P-R to ROC)}$$

Receiver Operator Characteristic (ROC) and Area Under the Curve (AUC)

- ROC curve takes FPR as x-axis, and TPR as y-axis.
- AUC is an area under the ROC curve



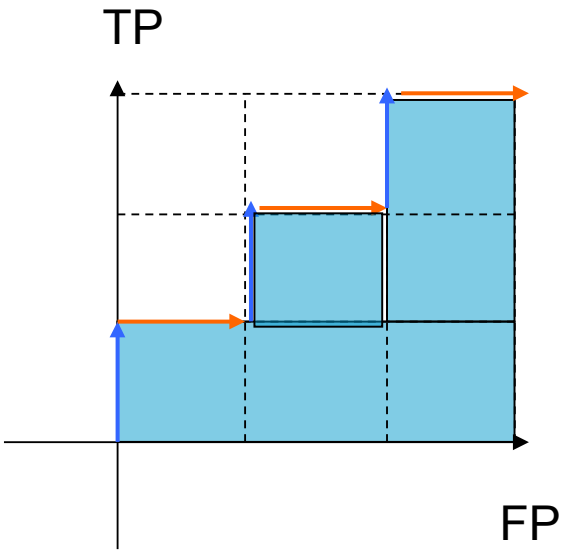
AUC = 6/9 = 0.66

(0 <= AUC <= 1)

Computing AUC

- Sort the prediction results in a descending order.
- By looking at score from the top one by one, then move rightward if FP, and upward if TP.
- Calculate the relative area under the ROC curve.

Predicted score	True Class	FP	TP
0.9	+	0	1
0.7	-	1	1
0.3	+	1	2
-0.1	-	2	2
-0.4	+	2	3
-0.6	-	3	3

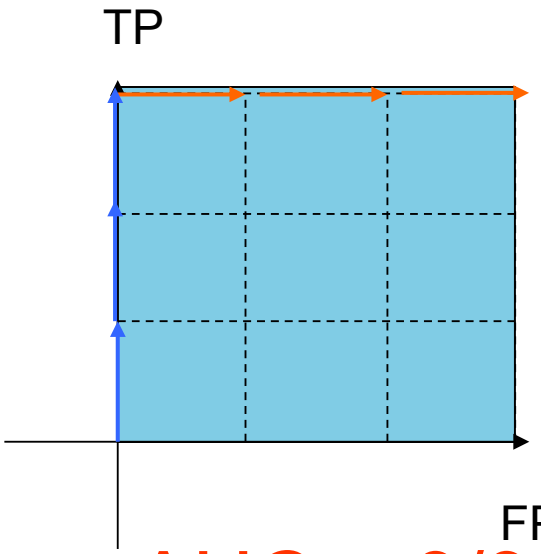


AUC = 6/9 = 0.66

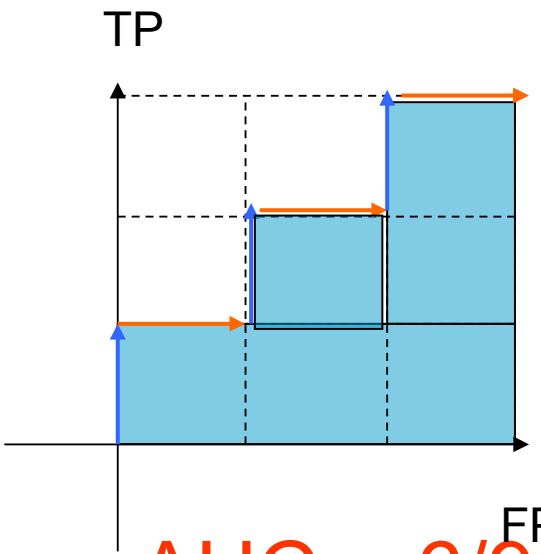
(0 <= AUC <= 1)

AUC(Area Under the Curve)

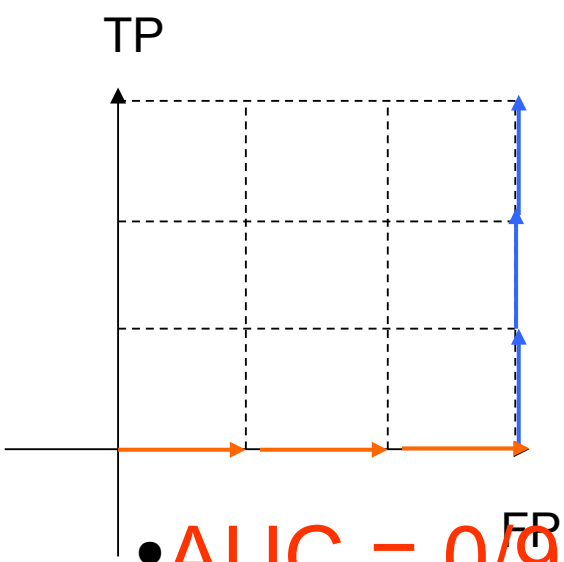
- 0<=AUC<=1
- AUC curve located above the other AUC curves suggests better performance.



•AUC = 9/9



•AUC = 6/9



•AUC = 0/9

Ex. 2

- Compute Accuracy, Sensitivity, Specificity, Recall, Precision, and FDR from the following table.

		actual	
		Yes(Class1)	No(Class2)
prediction	Yes(Class1)	300	15
	No(Class2)	50	1000

Ex. 3

- Calculate AUC score from the following table.

Prediction	True Class
1.0	+
0.8	+
0.5	+
0.3	-
0.1	+
-0.1	-
-0.2	+
-0.6	-
-0.9	-
-1.0	-

(Ex. 4) previous exercise revisited.

- Complete the following table by evaluating all the digits drawn by 200 people in test set T.
- Which digit is often misclassified ? To which digit ?

actual

prediction	1	2	3	4	5	6	7	8	9	0
1	199	0	0	0	1	0	0	0	0	0
2	0	169	8	8	1	2	4	8	0	0
3										
4										
5										
6										
7										
8										
9										
0										

Regarding personal project

- Analyze the data around you, and present it at the class.
- You can use publicly available dataset
  - http://archive.ics.uci.edu/ml/datasets.html
- 5 minutes presentation + 5 minutes discussion
- Select a problem which is not directly related to your bachelor/master thesis.