浙江大学 2013 －2014 学年 春季 学期

《Artificial Intelligence》课程期末考试试卷

课程号： 21190770 ，开课学院： 计算机科学与技术学院

考试试卷：A 卷、B 卷（请在选定项上打 √ ）

考试形式：闭、开卷（请在选定项上打 √ ），允许带_____入场

考试日期： 2014 年 4 月 22 日,考试时间： 120 分钟

诚信考试，沉着应考，杜绝违纪。

考生姓名：_____学号：_____所属院系：_____

题序	一	二	三	四	五	六	七	八	总 分
得分									
评卷人									

1. Fill in the blanks (30 points)

1) AI attempts not just to understand but also to build intelligent entities. Historically, there are many definitions of AI that can be organized into four categories: Thinking Rationally, Thinking Humanly, _____ and_____.

2) For a given probability distribution $p(\mathbf{x}|\mathbf{w})$, the conjugacy is that the _____ distribution has the same functional form as the _____ distribution.

3) Machine learning problems are usually divided into two main categories. Regression and classification are known as _____ learning problems. Clustering and density estimation are known as _____ learning problems.

4) The Linear basis function models involve linear combinations of fixed nonlinear functions of the input variables. If given basis functions $\Phi(\mathbf{x})=(1, \phi_1, \ldots, \phi_{M-1})^T$ and the model parameters $\mathbf{w}=(w_0, w_1, \ldots, w_{M-1})^T$, then the linear basis function $y(\mathbf{x}, \mathbf{w})$ = _____.

5) Given a training data set comprising N observations $\{\mathbf{x}_n\}$, together with corresponding target values $\{t_n\}$, we try to find a function $y(\mathbf{x}_n, \mathbf{w})$ to fit the train data by minimizing an error function, such as the sum-of-squares error (SSE) function which is defined by $E(\mathbf{w})$= _____.

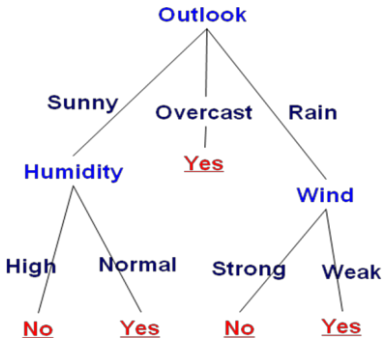
6) **Data points $\{\mathbf{x}_n\}$, $n = 1, \ldots, N$, that are drawn independently from the same Gaussian distribution $\mathrm{N}(\mathbf{x}_n \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ are said to be _____, thus the joint distribution $p(\mathbf{x}_1, \ldots, \mathbf{x}_N)$= _____.**

7) **A kernel function is a _____ function (symmetric/anti-symmetric) given by the relation $k(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x})^T \varphi(\mathbf{y})$. The simplest kernel function $k(\mathbf{x}, \mathbf{y})$ =_____.**

8) **Given a Gaussian Mixture Model $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathrm{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\mathbf{x}$ is a D-dimensional vector. In this model, each $\boldsymbol{\mu}_k$ contains D adjustable parameters, how many parameters should be estimated for the means, the covariance matrix and mixture coefficients totally? _____.**

9) **Many natural signals correspond to low-dimensional nonlinear manifolds embedded within the higher dimensional observed data space. Some manifold learning methods, such as _____, can model such nonlinear structure of data points and project the high-dimensional data points down to a lower dimensional subspace**

10) **Given a decision tree T shown on the following figure, if the input attributes are (Outlook=*Rain*, Temperature=*Cool*, Humidity=*High*, Wind=*Weak*), then the T will output_____.**



2. Multiple Choice (20 points)

1) **In 1950, Alan Turing proposed an approach, named _____, to test whether a computer would really be intelligent if a human interrogator, after posing some written questions, cannot tell whether the written responses come from a person or from a computer.**

A. Intelligent Agent **B. Turing Test**

C. Strong AI **D. Weak AI**

2) **In regression problems, we often need to minimize an error function that measures the misfit between the function output and the training set data points. For a given model, assume we have evaluated the sum-of-squares error $E_1(\mathbf{w})$ and $E_2(\mathbf{w})$ for two test data sets $\mathbf{D}_1$ and $\mathbf{D}_2$ with different size. We also computed the corresponding root-mean-square error $E_{\mathrm{rms1}}(\mathbf{w})$ and $E_{\mathrm{rms2}}(\mathbf{w})$. Which of the following discriminant condition can lead to the conclusion that this model has better fitting performance on test set $\mathbf{D}_1$? _____**

A. $E_1(\mathbf{w}) < E_{rms1}(\mathbf{w})$ **B.** $E_1(\mathbf{w}) < E_2(\mathbf{w})$

C. $E_{rms1}(\mathbf{w}) < E_{rms2}(\mathbf{w})$ **D.** $E_{rms1}(\mathbf{w}) < E_1(\mathbf{w})$

3) Assume D is the observed data set and w is model parameter. To determine the value of w, there are several methods. One is B **which maximizes $p(\mathbf{D}|\mathbf{w})$, one is _____ which maximize $p(\mathbf{w}|\mathbf{D})$ and the other minimizing an objective function such as loss function is called _____. (MAP—Maximum a Posterior, ML---Maximum Likelihood, LS—Least Squares)**

A. MAP, ML, LS. **B. ML, MAP, LS**

C. LS, ML, MAP **D. ML, LS, MAP**

4) Consider a polynomial curve fitting problem. If the fitted curve oscillates wildly through each point and achieve bad generalization by making accurate predictions for new data, we say this behavior is over-fitting. Which of the following methods cannot be used to control over-fitting? A

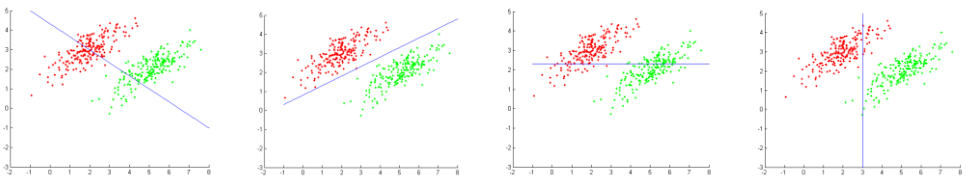
A. Use fewer training data

B. Add validation set, use Cross-validation

C. Add a regularization term to an error function

D. Use Bayesian approach with suitable prior

5) The following figures show some data points of two classes that are well separated in the original two-dimensional space. Now we consider using *Fisher*'s linear discriminant to project all data onto one-dimensional space and make prediction in the projected space. The line plotted on each figure represents the direction for projection of the data down to one dimension. Which of the following figure can give a larger separation between the projected class means while also giving a smaller variance within each class?_____



A. **B.** **C.** **D.**

6) Given two Gaussian distribution $N(x \mid 0, 1)$ **and** $N(x \mid 1, 1)$**, which of the following formula is correct?** _____

A. $N(0.5 \mid 0, 1) > N(0.5 \mid 1, 1)$ **B.** $N(1 \mid 0, 1) = N(0 \mid 1, 1)$

C. $N(0.5 \mid 0, 1) < N(0.5 \mid 1, 1)$ **D.** $N(0 \mid 0, 1) = N(0 \mid 1, 1)$

7) **To obtain an optimal value of w in linear regression, we can use stochastic gradient descent** $\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} - \eta \nabla E_n$, **where** $E(\mathbf{w}) = \frac{1}{2}(t_n - \mathbf{w}^T \phi_n)^2$, $\phi_n = \phi(\mathbf{x}_n)$. **Which of the following equation is correct?**

A. $\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} - \eta(t_n - \mathbf{w}^T \phi_n)\phi_n$

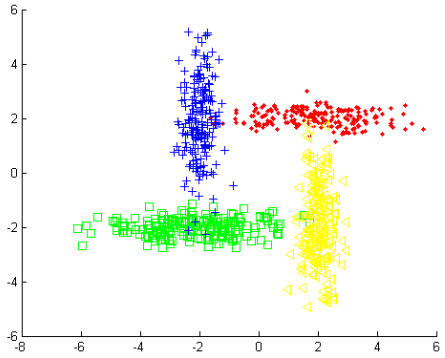
B. $\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} + \eta(t_n - \mathbf{w}^T \phi_n)\phi_n$

C. $\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} + \eta(t_n - \mathbf{w}^{\tau^T} \phi_n)\phi_n$

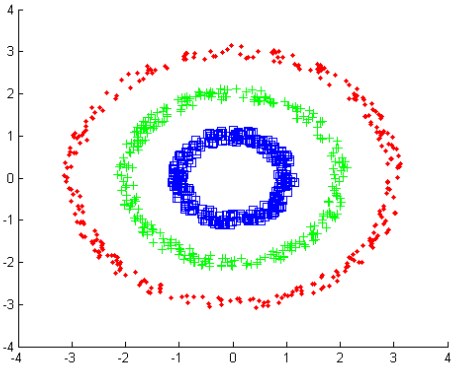
D. $\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} - \eta \phi_n{}^T (t_n - \mathbf{w}^{\tau^T} \phi_n)$

8) **The data points in the following figures (A to D) have different cluster structure in two-dimensional space. If we assume figure A, B, C and D have 4, 3, 4 and 3 clusters respectively (depicted in different marks) and apply k-means algorithm to find the clusters, which of the data set in the following figures can give the correct clustering result? _____**

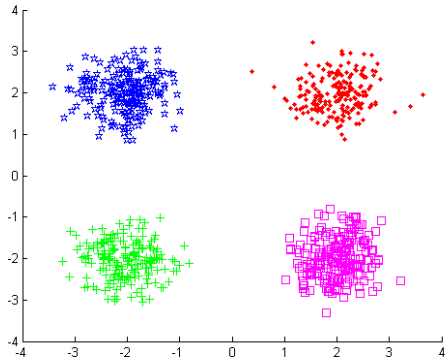
A.



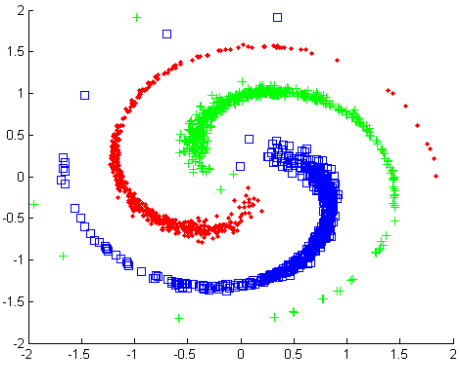
B.



C.

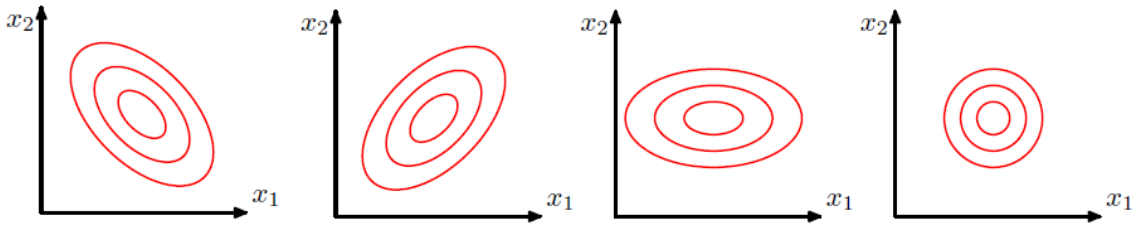


D.



9) **For a 2D multivariate Gaussian distribution** $N(\mathbf{x}|\, \mu, \Sigma)$, **if** $\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$, **which of the following figures is the contour of constant probability density of this Gaussian distribution?**

_____



A. **B.** **C.** **D.**

10) PCA can be defined as the orthogonal projection of the data onto a lower dimensional space, known as the principal subspace, such that the variance of the projected data is maximized. Which of the following statements about the applications of PCA is incorrect? _____

A. PCA can be used for dimensionality reduction.

B. PCA can be used for modeling nonlinear manifolds.

C. PCA can be used for data compression.

D. PCA can be used for data visualization.

3. Calculus, Analysis and Proof (50 points)

1) (10 points) Assume the error function with a regularization term in regression has given by $E(\mathbf{w}) = \frac{1}{2}\|\mathbf{t} - \mathbf{\Phi w}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$, where t is the target vector, λ is the regularization coefficient and $\mathbf{\Phi}$ is the design matrix. Find the solution of w by minimizing $E(\mathbf{w})$.

2) (15 points) Given a data set $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^T$ in which the observations $\{\mathbf{x}_n\}$ are assumed to be drawn independently from a multivariate Gaussian distribution given by Appendix 1.(b).

(a) Find the likelihood function $p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ (2 points)

(b) Find the log likelihood function $\ln p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ (4 points)

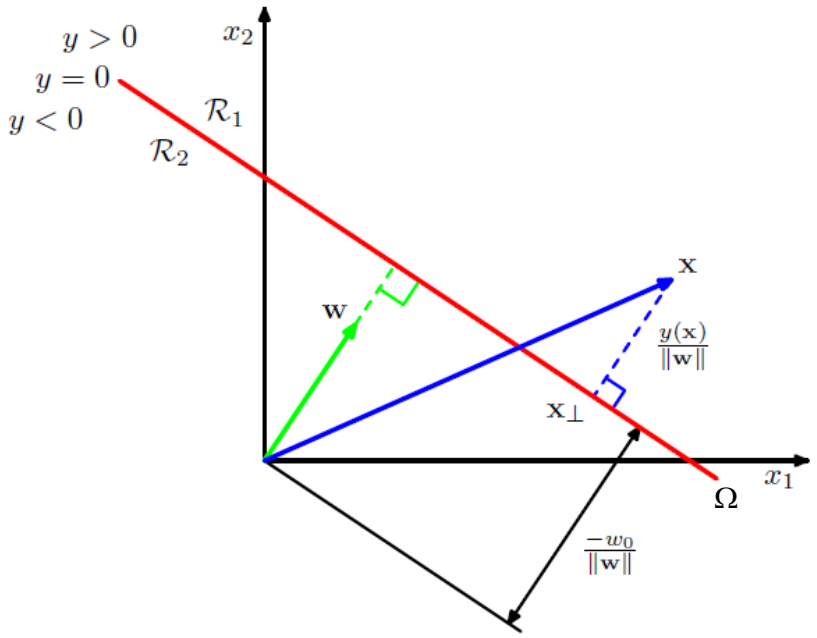
(c) Find the maximum likelihood solution of the mean $\boldsymbol{\mu}_{ML}$ (4 points)

(d) Find the maximum likelihood solution of the covariance $\boldsymbol{\Sigma}_{ML}$ (5 points)

3) (10 points) Consider a linear discriminant function $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ for two classes in 2-dimensional input space, the geometry is shown in the following figure. An input vector x is assigned to class C_1 if $y(\mathbf{x}) \geq 0$ and to class C_2 otherwise. The corresponding decision boundary is therefore defined by the relation $y(\mathbf{x}) = 0$, which corresponds to the line Ω in figure. Prove that:

(a) The vector w is orthogonal to every vector lying within the decision surface Ω. (5 points)

(b) The perpendicular distance r of arbitrary input vector x from the decision surface is $r = \frac{y(\mathbf{x})}{\|\mathbf{w}\|}$. (5 points)



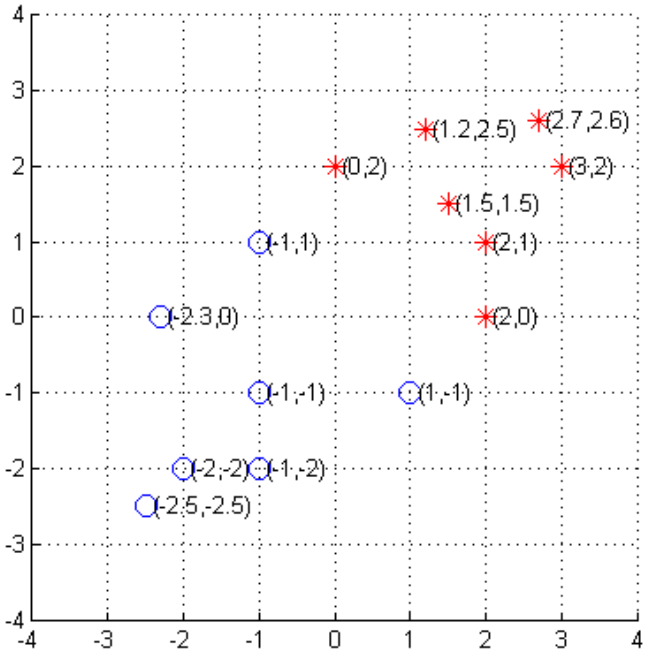
4) (15 points) This question includes two parts: (4A) and (4B). You just need to choose one of them to answer. If you answer both of them, then only the score of question (4A) is valid.

(4A) Assume we are using SVM approach to classify the two classes shown below, with '✳' denoting one class and '○' another.

(a) Find and list the corresponding support vectors. (5 points)

(b) Use the result of (a) to find the equation of decision surface and plot the final decision surface in the figure. (5 points)

(c). Use the result of (b), draw the margin of SVM in the figure, then give its exact value. (5 points)



(4B) Compare the difference between two main machine learning problems: supervised learning and unsupervised learning (Please give some detailed description of regression, classification, clustering and so on.) (15 points).

Appendix:

1. Probability distributions:

（a） **Single variable Gaussian:** $N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$

（b） **D-dimensional multivariate Gaussian:**

$$N(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\},$$

$$-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + const$$

（c） **Beta:** $Beta(\mu \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$

（d） **Dirichlet:**

$$Dir(\boldsymbol{\mu} \mid \boldsymbol{\alpha}) = C(\boldsymbol{\alpha}) \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}, \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix}, \boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{pmatrix}, 0 \le \mu_k \le 1, \sum_{k=1}^{K} \mu_k = 1$$

（e） **Gamma:** $Gam(\tau \mid a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau}, a > 0, b > 0, \tau > 0$

2. Matrix calculus

（a） $\left(\mathbf{AB}\right)^T = \mathbf{B}^T\mathbf{A}^T, \quad \mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}, \quad \left(\mathbf{AB}\right)^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

（b） $Tr\left(\mathbf{AB}\right) = Tr\left(\mathbf{BA}\right), \quad Tr\left(\mathbf{ABC}\right) = Tr\left(\mathbf{CAB}\right) = Tr\left(\mathbf{BCA}\right), \quad \left|\mathbf{AB}\right| = \left|\mathbf{A}\right|\left|\mathbf{B}\right|$

（c） $\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix}, \mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}$

（d） $\frac{\partial(\mathbf{x}^T\mathbf{a})}{\partial\mathbf{x}} = \frac{\partial(\mathbf{a}^T\mathbf{x})}{\partial\mathbf{x}} = \mathbf{a} \qquad \frac{\partial\mathbf{a}^T\mathbf{X}^{-1}\mathbf{b}}{\partial\mathbf{X}} = -\mathbf{X}^{-T}\mathbf{a}\mathbf{b}^T\mathbf{X}^{-T} \qquad \frac{\partial\mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial\mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$

（e） Assume $\boldsymbol{\Lambda}$ is symmetric matrix, then $\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu}) = 2\boldsymbol{\Lambda}(\mathbf{x}-\boldsymbol{\mu})$

（f） $\frac{\partial}{\partial\mathbf{w}}(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w})^T(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}) = -2\boldsymbol{\Phi}^T(\mathbf{t}-\boldsymbol{\Phi}\mathbf{w}), \quad \frac{\partial}{\partial\mathbf{w}}\left\|\mathbf{w}\right\|^2 = \frac{\partial}{\partial\mathbf{w}}\mathbf{w}^T\mathbf{w} = 2\mathbf{w}$

（g） $\frac{\partial}{\partial\mathbf{W}} Tr\left[(\mathbf{T}-\boldsymbol{\Phi}\mathbf{W})(\mathbf{T}-\boldsymbol{\Phi}\mathbf{W})^T\right] = -2\boldsymbol{\Phi}^T(\mathbf{T}-\boldsymbol{\Phi}\mathbf{W})$

（h） $\frac{\partial}{\partial\mathbf{A}}\ln\left|\mathbf{A}\right| = (\mathbf{A}^{-1})^T, \quad$ if $\mathbf{A} = diag(\lambda_i),\ i = 1,...,D,$ then $\left|\mathbf{A}\right| = \prod_{i=1}^{D}\lambda_i$

《Artificial Intelligence》

Final Examination Answer Sheet

Name: _____Student ID: _____Dept.: _____

Section	1	2	3	Total
Score				
Reviewer				

1. Fill in the blanks (30 points, 2pt/per)

1)._____, _____

2)._____, _____

3)._____, _____

4)._____

5)._____

6)._____, _____

7)._____, _____

8)._____

9)._____

10)._____

2. Multiple Choice (20 points, 2pt/per)

1	2	3	4	5	6	7	8	9	10

3. Calculus, Analysis and Proof (50 points)

1) (10 points)

2) (15 points)

(a) (2 points)

(b) (4 points)

(c) (4 points)

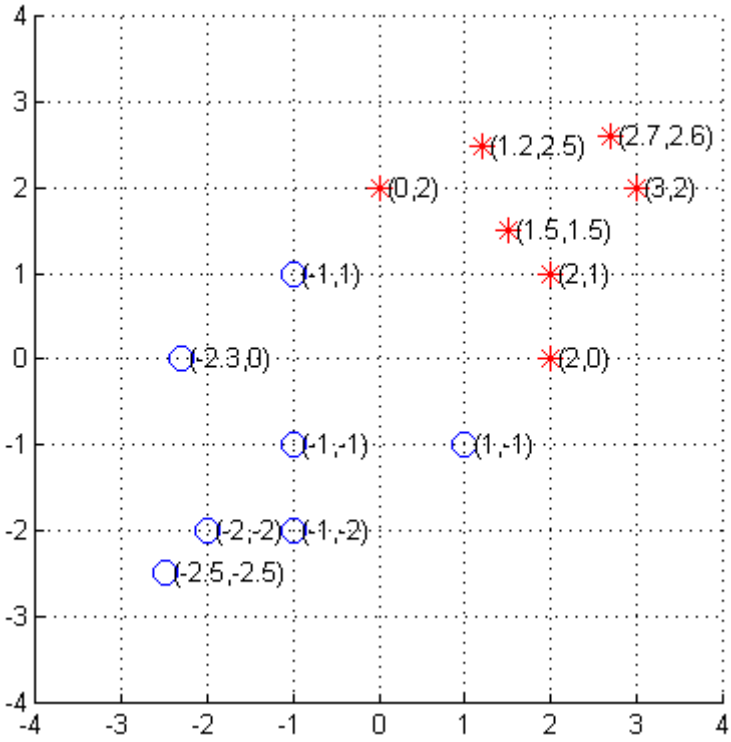
(d) (5 points)

3) (10 points)
(a) (5 points)

(b) (5 points)

4) (15 points) This question includes two parts: (4A) and (4B). You just need to choose one of them to answer. If you answer both of them, then only the score of question (4A) is valid.

（4A）
(a) (5 points)



(b) (5 points)

(c) (5 points)

（4B） (15 points)