浙江大学 2012 －2013 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： 2013 年 4 月 29 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 Humanly, __________, ___________ and Acting Rationally.

2) 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.

3) The assertion that machines could act as if they were intelligent is called the _________ hypothesis by philosophers, and assertion that machines that do so are actually thinking (not just simulating thinking) is called the ___________ hypothesis.

4) Machine learning is usually divided into two main types. Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as ____________ learning problem. Applications in which the training data consists of a set of input vectors without any corresponding target values are known as ____________ learning problem.

5) **Assume the sum-of-squares error function is defined by E(w), then the root-mean-square error function can be defined by ___________________.**

6) **Data points that are drawn independently from the same distribution are said to be ___________________, which is often abbreviated to _____________.**

7) **Given multivariate Gaussian distribution $N(x|\mu, \Sigma)$, then Mahalanobis distance Δ is defined by _______________________.**

8) **For a given probability distribution p(x|w), if we choose a prior p(w), then the posterior distribution p(w|x) will have the same functional form as the prior. This property is called _____________ .**

9) **In generalized linear models, if activation function is logistic sigmoid function f(.), then the corresponding link function is ____________.**

10) **The generalized linear model based on a probit activation function is known as ________________. The inverse probit function can be constructed by the _______________ function of a zero mean, unit variance Gaussian N(x|0,1).**

2. Multiple Choice (20 points)

1) **Which of the following statements about ML problems is false? ________**

A. The regression is one of unsupervised learning problems.

B. The classification is one of supervised learning problems.

C. The clustering is one of unsupervised learning problems.

D. The density estimation is one of unsupervised learning problems.

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 $E_1(w)$ and $E_2(w)$ for two test data sets D_1 and D_2 with different size. We also computed the corresponding root-mean-square error $E_{rms1}(w)$ and $E_{rms2}(w)$. Which of the following discriminant condition can lead to the conclusion that this model has better fitting performance on test set D_1? ___________**

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

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

3) 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. 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

4) Assume D is the observed data set and w is model parameter. Which of the following statements about likelihood function p(D|w) is false? _____

A. It expresses how probable the observed data set is for different settings of the parameter vector w.

B. The likelihood is not a probability distribution over w.

C. Its integral with respect to w must be equal to one.

D. Maximizing the likelihood function is equivalent to minimizing the error.

5) Which of the following statements about the Fisher's criterion is correct? __________

A. It maximizes the separation between the projected class means as well as the total within-class variance.

B. It minimizes the separation between the projected class means as well as the total within-class variance.

C. It maximizes the separation between the projected class means as well as the inverse of the total within-class variance.

D. It minimizes the separation between the projected class means as well as the inverse of the total within-class variance.

6) Given two Gaussian distribution N(x|0,1) and N(x|1,1), which of the following formula is correct? ______________

A. N(0.5|0,1) > N(0.5|1,1) B. N(1|0,1) = N(0|1,1)

C. N(0.5|0,1) < N(0.5|1,1) D. N(0|0,1) = N(0|1,1)

7) Which of the following statements about the kernel function is false? _________

A. The kernel function is a symmetric function.

B. The simplest example of a kernel function is $k(x, x') = x^T x'$.

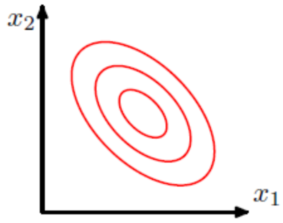
C. The feature vector that corresponds to the Gaussian kernel has infinite dimensionality.

D. We cannot construct new kernels by using simpler kernels.

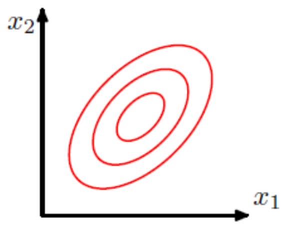
8) Assume the precision matrix is given by $\mathbf{R}=\begin{pmatrix}\boldsymbol{\Lambda}+\mathbf{A}^T\mathbf{L}\mathbf{A} & -\mathbf{A}^T\mathbf{L}\\ -\mathbf{L}\mathbf{A} & \mathbf{L}\end{pmatrix}$, then the corresponding covariance matrix $\boldsymbol{\Sigma}=\mathbf{R}^{-1}=\begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Lambda}^{-1}\mathbf{A}^T\\ \mathbf{A}\boldsymbol{\Lambda}^{-1} & \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T\end{pmatrix}$, find the expression of $\boldsymbol{\Sigma}_{11}=$___________.

A. $\mathbf{L}^{-1}$ B. $\boldsymbol{\Lambda}^{-1}$ C. $\mathbf{A}^T$ D. $\mathbf{A}^{-1}$

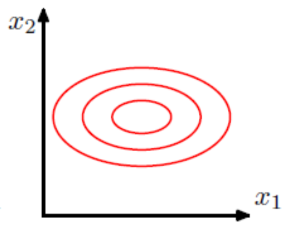
9) For a 2D multivariate Gaussian distribution N(x| μ , Σ), if $\Sigma=\begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}, \sigma_1^2>\sigma_2^2$, which of the following figures is the contour of constant probability density of this Gaussian distribution? _______________



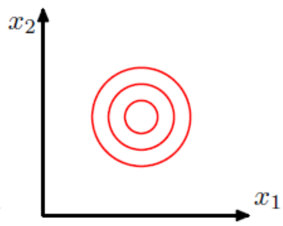
A.



B.



C.



D.

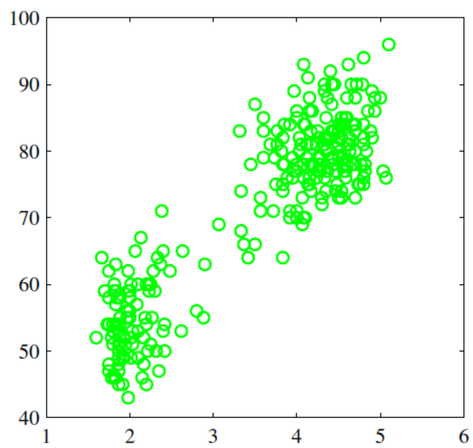
10) For the 'Old Faithful' data shown on the following figure, which probabilistic model can represent it more accurately? _______

A. Gaussian distribution

B. Mixture of Gaussians

C. Dirichlet distribution

D. Wishart distribution



3. Calculus, Analysis and Proof (50 points)

1) Consider the multivariate Gaussian distribution given by Appendix 1.(b). By writing the precision matrix (inverse covariance matrix) $\mathbf{\Lambda} = \mathbf{\Sigma}^{-1}$ as the sum of symmetric matrix $\mathbf{S} = (\mathbf{\Lambda} + \mathbf{\Lambda}^T)/2$ and anti-symmetric matrix $\mathbf{A} = (\mathbf{\Lambda} - \mathbf{\Lambda}^T)/2$, show that:

(a) the inverse matrix $\mathbf{S}^{-1}$ is symmetric. (3 points)

(b) the anti-symmetric term A does not appear in the exponent of the Gaussian for $\mathbf{\Lambda} = \mathbf{S} + \mathbf{A}$, such that $(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{S}(\mathbf{x}-\boldsymbol{\mu})$. (4 points)

2) Assume an eigenvalue decomposition of the covariance matrix $\mathbf{\Sigma}$ is given by $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$, where $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_D), \mathbf{U}^T\mathbf{U} = \mathbf{I}, \mathbf{\Lambda} = diag(\lambda_i), i = 1, \ldots, D$, show that:

(a) $\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^T$ (4 points)

(b) $|\mathbf{\Sigma}| = \prod_{i=1}^{D} \lambda_i$ (4 points)

3) Given a data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$ 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 $p(\mathbf{X} \mid \boldsymbol{\mu}, \mathbf{\Sigma})$ (2 points)

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

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

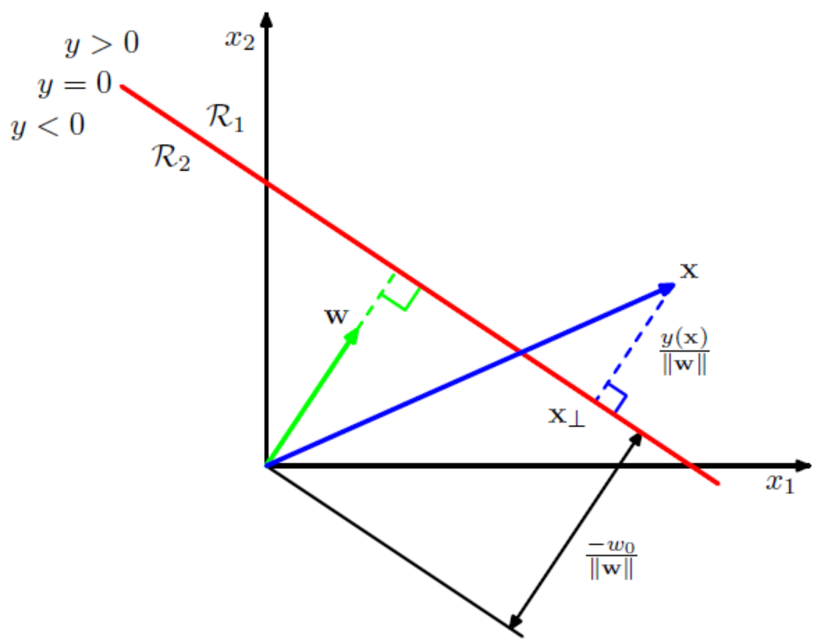
(d) Find the maximum likelihood solution of the covariance $\boldsymbol{\Sigma}_{ML}$ (5 points)

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

5) **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 $r = \dfrac{y(\mathbf{x})}{\|\mathbf{w}\|}$. (5 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:**

$$Dir(\boldsymbol{\mu}\mid\boldsymbol{\alpha})=C(\boldsymbol{\alpha})\prod_{k=1}^{K}\mu_k^{\alpha_k-1},\boldsymbol{\mu}=\begin{pmatrix}\mu_1\\ \vdots\\ \mu_K\end{pmatrix},\boldsymbol{\alpha}=\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_K\end{pmatrix},0\le\mu_k\le1,\sum_{k=1}^{K}\mu_k=1$$

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

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

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

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

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

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

《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) (7 points)

(a) (3 points)

(b) (4 points)

2) (8 points)

(a) (4 points)

(b) (4 points)

3) (15 points)

(a) (2 points)

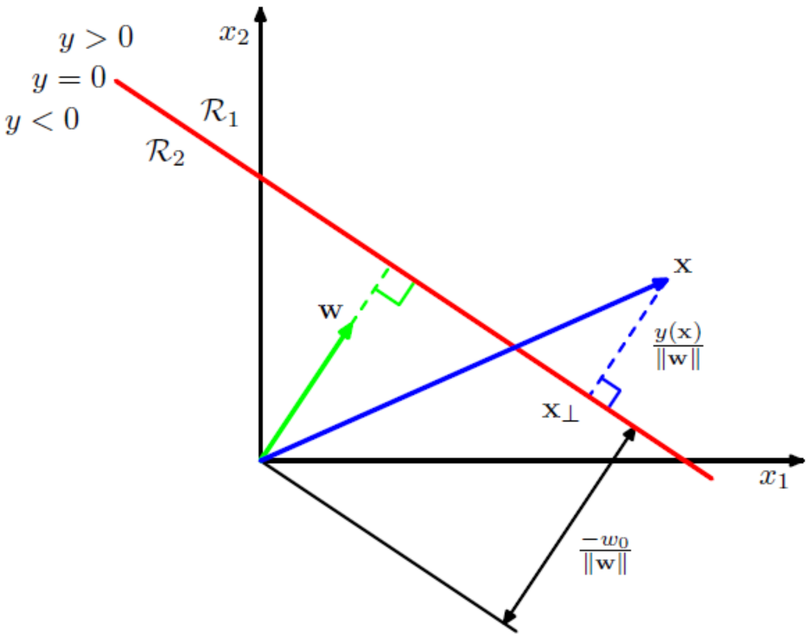
(b) (4 points)

(c) (4 points)

(d) (5 points)

4) (10 points)

5) (10 points)



(a) (5 points)

(b) (5 points)