CS 725 Foundations of Machine Learning
Endsemester Examination; **2pm-5pm**, 28-Apr-2015

Roll No. ______________________

Important Instructors

- Please fill in your **roll number** in the blank above.
- Please fill in the blanks with appropriate answers. There is **no separate answer sheet**. Answers written beyond the space provided may NOT be evaluated.
- It is recommended that you first write the solutions and answers in rough sheets. This will also help you to understand whether your answer will fit the space provided or you will need to shorten it by skipping less important details etc. Once you are convinced that the answer cannot be further improved, then only make a fair version of it in the space provided for the answer. You may take as many rough sheets as you need.
- Note that the markings will be binary based on the exactness of your answers written in the blanks. So make sure your answers are legible, precise, and do not commit silly mistakes/typos in writing your answers.
- Your answers must be in a **highly simplified** form. For e.g., if the correct answer is $\int_a^b x^2 \, \mathrm{dx} \geq 0$, then you need to write the answer as $b \geq a$. Answers that are not simplified will NOT get any credit.
- Needless to say, a good way of verifying your answer is to check if it intuitively, geometrically, syntactically makes sense.
- This is a closed book exam. Electronic gadgets, including a calculator, are strictly NOT allowed.

- All questions are of fill-in-the-blank type. To make the questions precise, the type of answer to be filled in the blank is clearly mentioned after the blank. For example, if a blank is followed by:
  - [[MathExpr]], then you need to write an answer that is a valid mathematical expression. For e.g., $3x - 4y + x^3$, (or) $\neq 0$, (or) $trace(M)$ etc.
  - [[Numeric]], then you need to write an answer that is a valid number. For e.g., 3/4, (or) -3.14, (or) 5, (or) irrationals like ϕ, e. etc.
  - [[Term]], then you need to write an answer that is an English phrase representing a well-defined object/concept or a well-known theorem name. For e.g., equilateral triangle (or) Spectral theorem etc.
  - [[T/F/M]], then the only choices for answer are 'T', 'F', and 'M'. While 'T', 'F' represents that the previous sentence is 'true', 'false' respectively; 'M' represents that the given information is in-sufficient to decide whether the previous sentence is true or false (basically 'M' handles the undecidable case). If [[T/F]] is given, then the only choices for answer are 'T', and 'F'.
  - a list of semi-colon separated choices, then the answer MUST be one of those choices alone. For e.g., If [[e^x ; $\log(x)$]] is given, then the only choices for answer are 'e^x', and '$\log(x)$'. If [[1 ; 2 ; $\ldots$; n]] is given, then the only choices for answer are '1', '2' and so on upto n.
  - [[Reason]], then you need to write a short justification for the answer you wrote in the blank preceeding this.
- Sometimes the type is not designated for each blank, but for a group of blanks. For e.g. four blanks may be followed by the code [[Numeric]], then the interpretation is that each of the four blanks are to be filled with a numeric value.
- In case your answer for a blank is wrong, but is of the correct type (as per the code provided), then you get a zero for that blank. However, in case your answer is of wrong type (as per the code), then you get negative (-) of half of the marks allocated for the blank, even if your answer "otherwise is correct".

Section 1. Fill in the blanks

1. Let $\mathcal{D} = \{(x_1, y_1), \ldots, (x_9, y_9)\}$ be a given training set, where each $x_i \in \mathcal{X}$, $y_i \in \{-1, 1\}$. Let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be a valid kernel. Let $\phi : \mathcal{X} \mapsto \mathbb{R}^n$ be such that $k(x, z) = \phi(x)^\top \phi(z)$, $\forall\ x, z \in \mathcal{X}$. Consider a 1-NN classifier[1] built in the R^n space induced by k. i.e., label of x is assigned as y_i if $\phi(x_i)$ is closest to $\phi(x)$ among all $\{\phi(x_1), \ldots, \phi(x_9)\}$. If the kernel values are as given in the table below, then the label of x is ___[[y_1 ; y_2 ; ... ; y_9]]. (1mark)

$k(x_1,x)$	$k(x_2,x)$	$k(x_3,x)$	$k(x,x_4)$	$k(x,x_5)$	$k(x,x_6)$	$k(x_7,x)$	$k(x,x_8)$	$k(x_9,x)$
3	1.1	0.99	45	45.1	44.9	0	0.1	0.1

The label of x according to 3-NN (with majority vote) is ______________________[[MathExpr]]. (1mark)

In the context of this application, Mr. Satyam claims that the true probabilities of misclassification with 1-NN, 3-NN asymptotically[2] converge to 0.3 and 0.1 respectively[3]. He is lying.
___[[T/F/M]].____[[Reason]]
(2marks)

2. Consider the following dataset[4]:

$\phi_1(x_i)$	5	3	0	1	0	5	10	1	-5	-10	-1	-2	-15	-1	-5	-1
$\phi_2(x_i)$	3	5	3	10	-3	2	-5	2	0	-4	0	-3	-10	-5	-4	-6
y_i	1	1	-1	1	-1	-1	-1	1	-1	-1	-1	1	1	1	1	-1

Apply the decision tree algorithm taught in lectures[5] on the above data. The root node of the decision tree hence constructed will[6] be "$\phi_1(x) \geq$__[[Numeric]]".(2marks). The nodes corresponding to the true and false evaluation of the root node will be "$\phi_2(x) \geq$__[[Numeric]]"

---

[1] NN stands for Nearest Neighbour.

[2] As the no. of training instances goes to ∞.

[3] Mr. Satyam has access to the true (but unknown to us) distribution.

[4] First row gives the value of the first feature for every training instance. The next shows the value of the second feature; whereas the last one shows the labels.

[5] Assume that the first feature is evaluated at level one and the next at the subsequent level.

[6] There are multiple "correct answers". Hence it is enough to write any one of them.

and "$\phi_2(x) \geq$__[[Numeric]]" respectively.(2marks). Finally, the labels of the two most extreme leaves i.e., true evaluation at both levels and false evaluation at both levels will be _[[-1 ; 1]] and _[[-1 ; 1]] respectively. (1mark)

3. Consider a dataset with 2 points in 1d: $(x_1 = 0, y_1 = -1)$ and $\left(x_2 = \sqrt{2}, y_2 = 1\right)$. A hard-margin SVM[7] (with the threshold term b that is not fixed to zero) is constructed using the quadratic kernel given by $k(x, z) = (1 + xz)^2$. Then at optimality $w = [\_\_\_;\ \_\_\_;\ \_\_\_]^\top$ and $b =$ __[[Numeric]] respectively. (2marks).

4. The following table provides the values of $w^\top \phi(x_i), y_i$ and α_i, where w is the optimal w obtained by training an SVM[8] on a training set and $\phi(x_i), y_i, \alpha_i$ denote the feature vector, label and dual variable respectively for the i^{th} training instance.

$w^\top \phi(x_i)$	95	-98	97	3	-7	-1	23	20	-37	-54	62	91	-55	23	-1	1
y_i	-1	1	-1	1	-1	1	-1	-1	1	1	1	-1	-1	1	1	1
α_i	C	C	C	0	0	C	C	C	C	C	0	C	0	0	C	0

Then, b is optimal if and only if b lies in the interval [__, _][[Numeric]][9] (2marks).

5. Ridge regression is not an example of a support vector method.__[[T/F/M]].(0.5marks). The optimization problem arising in ridge regression also arises in MLE estimation for discriminative modeling of the regression problem.__[[T/F/M]]. (0.5marks)

6. Let $\mathcal{Y}$ be some set and let $\bar{\mathcal{X}}, \bar{\mathcal{Y}}$ denote the set of all sequences (of finite length) with elements from $\mathcal{X} \equiv \mathbb{R}^2, \mathcal{Y}$ repectively. Consider the HMM model with the emission distribution modelled by a (multivariate) Gaussian: $f_X(x) \propto \exp\left\{w^\top \phi(x) - A(w)\right\}$, where $x \in \mathbb{R}^2$ and w is the parameter vector for the Gaussian. If $x = [x_1; x_2]^\top$,

[7] Hard-margin SVM is the extreme case where absolutely no errors are allowed i.e., no loss is incurred at any training points. In other words, all the slack variables are at zero.

[8] SVM here means the usual hinge-loss based one and the threshold term b NOT fixed to zero. The regularization hyperparameter is C. Hence every $\alpha_i \in [0, C]$.

[9] Needless to say, in case b can take a unique value, say 3, then you can write it as the following interval $[3, 3]$.

then $\phi(x) = [\underline{\qquad\qquad\qquad\qquad}]^\top$[[MathExpr]](1mark). A struct-SVM version analogous to this HMM was taught in lectures. For that particular $\bar{\phi} : \bar{\mathcal{X}} \times \bar{\mathcal{Y}} \mapsto \mathbb{R}$, write down the equivalent kernel $k : \mathcal{XY} \times \mathcal{XY}$, where $\mathcal{XY} \equiv \bar{\mathcal{X}} \times \bar{\mathcal{Y}}$, in terms of the following two kernels: $k_1 : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ and $k_2 : \mathcal{Y} \times \mathcal{Y} \mapsto \{0, 1\}$ defined by $k_1(x, z) \equiv \phi(x)^\top \phi(z), \ \forall \ x, z \in \mathcal{X}$ and $k_2(y, u) \equiv \begin{cases} 1 & \text{if } y = u, \\ 0 & \text{if } y \neq u \end{cases}, \ \forall \ y, u \in \mathcal{Y}$. $k\left((x_1, y_1), (x_2, y_2)\right) \equiv$ _[[MathExpr]](2marks). In filling this blank please use the following notation: $x_i = (x_{i1}, \ldots, x_{iT_i}), y_i = (y_{i1}, \ldots, y_{iT_i})$, where all $x_{ij} \in \mathcal{X}$ and $y_{ij} \in \mathcal{Y}$.