SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

50.007 Machine Learning
2015 Term 6

Date: 6th Nov 2015
Time: 2:30 PM
Duration: 2 hours

Instructions to Candidates:

1. This paper consists of 11 questions with 5 printed pages (This title page counts as the first page).
2. This is a closed book examination.
3. Cheet sheets are not allowed.
4. Answer all the questions.
5. Write your answers in the answer books provided.
6. Wish you success!

1. (2P) Multiple Choice: Ridge regression optimizes what objective function?

   O mean average precision with squared ℓ_2-norm $\|w\|_2^2 = \sum_{d=1}^{D} w_d^2$ on weights w

   O mean squared error with squared ℓ_2-norm $\|w\|_2^2 = \sum_{d=1}^{D} w_d^2$ on weights w

   O mean squared error with ℓ_1-norm $\|w\|_1 = \sum_{d=1}^{D} |w_d|$ on weights w

2. (2P) Multiple Choice: The Support vector machine optimizes what objective function?

   O mean zero-one error with squared ℓ_2-norm $\|w\|_2^2 = \sum_{d=1}^{D} w_d^2$ on weights w

   O mean squared error with squared ℓ_2-norm $\|w\|_2^2 = \sum_{d=1}^{D} w_d^2$ on weights w

   O mean hinge loss with squared ℓ_2-norm $\|w\|_2^2 = \sum_{d=1}^{D} w_d^2$ on weights w

3. (1P) Multiple Choice: Neural Networks require feature mappings $\phi : x \mapsto \phi(x)$ in the lower layers to be designed by hand?

   O Yes

   O No

4. (2P) Suppose you have two data samples $x_1, x_2 \in \mathbb{R}^3$ with

$$x_i = \begin{pmatrix} x_i^{(1)} \\ x_i^{(2)} \\ x_i^{(3)} \end{pmatrix},$$

   then write down a matrix transformation $A \in \mathbb{R}^{3\times 3}$ such that the distance $\|Ax_1 - Ax_2\|$ between sample x_1 and x_2 is equal to

$$\|Ax_1 - Ax_2\| = \sqrt{(x_1^{(1)} - x_2^{(1)})^2 + 4(x_1^{(2)} - x_2^{(2)})^2}$$

5. (4P)

   - given a set of $((f(x_1), y_1), \ldots, (f(x_N), y_N))$ consisting predictions $f(x_i)$ on features x_i and their corresponding ground truth labels y_i, write down the formula for hinge loss averaged over the number of samples N

- given a set of $((f(x_1), y_1), \ldots, (f(x_N), y_N))$ consisting predictions $f(x_i)$ on features x_i and their corresponding ground truth labels y_i, write down the formula for mean squared error (MSE) averaged over the number of samples N

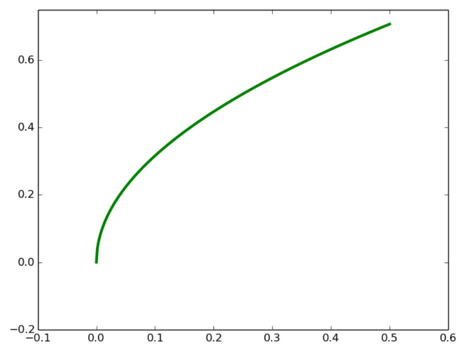
6. (4P) compute some derivatives with respect to vector x: either compute the gradient vector

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix}$$

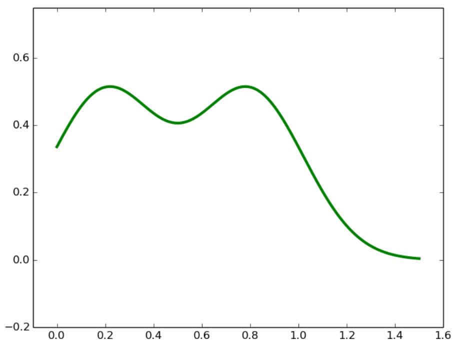
   or the linear mapping $Df(x)[h]$ which maps a direction vector h onto the directional derivative $Df(x)[h]$ of function f at point x. You can choose for each task separately whether you want to compute the gradient vector or the linear mapping. Hint: think where to apply product rule, and chain rule. Note that product rule is compatible with matrix multiplication, no matter whether the component is a matrix or a vector.

   (a) $f(x) = 3x^7 - 5x^5 + 7x^2 - 3$, $x \in \mathbb{R}^1$
   (b) $f(x) = x^\top Az \sum_{r=1}^{d} x^{(r)}$, $x \in \mathbb{R}^{d\times 1}, A \in \mathbb{R}^{d\times d}, z \in \mathbb{R}^{d\times 1}$, where $x^{(r)}$ is the r-th dimension of vector x
   (c) $f(x) = (x^\top Ax)^4$, $x \in \mathbb{R}^{d\times 1}, A \in \mathbb{R}^{d\times d}$
   (d) $f(x) = \|Ax\|_2\, x^\top x$, $x \in \mathbb{R}^{d\times 1}, A \in \mathbb{R}^{d\times d}$

7. (1P) is this function convex, concave or none of both?



8. (1P) is this function convex, concave or none of both?



9. (2P) A task for thinking a bit: Why is the following measure no good objective function for measuring the error in a regression problem ? The error is computed between ground truth y_i and prediction $f(x_i)$ as given by the function

$$E = \frac{1}{N}\sum_{i=1}^{N}(f(x_i) - y_i)^3$$

   Hint: you can imagine what can happen if this objective is used with a linear model: $f(x_i) = w^\top x_i$.

10. (3P) A more complicated task: Consider the the following distribution function (a special case of a so-called gamma distribution). It is defined for positive real numbers $x > 0$.

$$p(x) = cx^2 \exp(-\beta x)\beta^3$$

   Note x, c and β are real numbers. $c > 0, \beta > 0$.

   Your task: compute the maximum likelihood estimator for parameter β for given data $x_1, \ldots, x_n$. We assume that the samples x_i are independent.

   Hint: compute the maximum likelihood estimator as the maximum of the log-likelihood, that means after applying a logarithm.

11. (3P) Another somewhat more complicated task: which of the following functions is a kernel and which is not ? Give an argument why.

Suppose that $x_i = \begin{pmatrix} x_i^{(1)} \\ x_i^{(2)} \end{pmatrix}$ are two-dimensional vectors.

$$k(x_1, x_2) = x_1^\top \begin{pmatrix} 25 & 0 \\ 0 & 9 \end{pmatrix} x_2$$
$$k(x_1, x_2) = x_1^\top \begin{pmatrix} 25 & 0 \\ 0 & -9 \end{pmatrix} x_2$$
$$k(x_1, x_2) = x_1^\top \begin{pmatrix} 2 & 0 \\ 1 & 2 \end{pmatrix} x_2$$

Hint: Recall: if $k(x_1, x_2)$ is a kernel function, then there must exist a mapping ϕ and a Hilbert space such that

$$k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$$

where

$$\langle v_1, v_2 \rangle$$

is the inner product in the Hilbert space for two vectors v_1 and v_2, and

$$k(x_1, x_1) = \|\phi(x_1)\|^2$$

is the norm of $\phi(x_1)$ in this Hilbert space. Check whether you can find an explicit mapping ϕ for the euclidean inner product, or whether you can find a contradictions to properties of a kernel.

End of Paper