UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE : **COMPGI13**

ASSESSMENT PATTERN : **COMPGI13B**

MODULE NAME : **Advanced Topics in Machine Learning**

DATE : **01-Jun-12**

TIME : **10:00**

TIME ALLOWED : **2 Hours 30 Minutes**

2011/12-COMPGI13B-001-EXAM-17

TURN OVER

Advanced Topics in Machine Learning, GI13, 2011/12

Answer any THREE questions. Each question is worth 20 marks. Use separate answer books for PART A and PART B. **Gatsby PhD students only**: answer *either* TWO questions from PART A and ONE question from PART B; *or* ONE question from PART A and TWO questions from PART B.

Marks for each part of each question are indicated in square brackets

Calculators are NOT permitted

Part A: Kernel Methods

1. • Given the Gaussian kernel on $\mathbb{R}$,

$$k(x_i,x_j) = \kappa(x_i - y_j) = \exp\left(\frac{-1}{2}\left(x_i - x_j\right)^2\right), \tag{1}$$

we may define an RKHS with inner product

$$\langle f,g\rangle_{\mathcal{H}} = (2\pi^{-1/2})\int_{\mathbb{R}} \frac{\hat{f}(\omega)\overline{\hat{g}(\omega)}}{\hat{\kappa}(\omega)}d\omega, \tag{2}$$

where $\hat{f}$ is the Fourier transform of f,

$$f(x) = (2\pi)^{-1/2}\int_{\mathbb{R}} \hat{f}(\omega)e^{\iota\omega^\top x}d\omega.$$

What is the narrowest Gaussian that is in this RKHS? i.e., given a function

$$f(x) = \exp\left(-ax^2\right)$$

where $a > 0$, what is the minimum a for which $\|f\|_{\mathcal{H}} < \infty$? You will need: the Gaussian $g(x) = e^{-x^2/2}$ has a Fourier transform which is also a Gaussian, $\hat{g}(\omega) = e^{-\omega^2/2}$, and $f(ax)$ has Fourier transform $a\hat{f}(\omega/a)$.

[4 marks]

• Define the Laplace kernel on $\mathbb{R}$,

$$k_l(x_i,x_j) = \kappa_l(x_i - x_j) = \exp\left(\frac{-1}{2}\left|x_i - x_j\right|\right),$$

with Fourier transform

$$\hat{\kappa}_l(\omega) = \frac{2}{1+\omega^2}.$$

Given the RKHS inner product in eq. (2), comment on the smoothing penalty enforced by the RKHS norm $\|f\|_{\mathcal{H}}$ for the Gaussian kernel in eq. (1), vs that with the Laplace kernel.

[4 marks]

- The solution to the ridge regression problem,

$$f^* = \arg\min_{f\in\mathcal{H}} \left(\sum_{i=1}^{n} (y_i - \langle f, \phi(x_i)\rangle_{\mathcal{H}})^2 + \lambda \|f\|^2_{\mathcal{H}} \right);$$

takes the form

$$f^* = \sum_{i=1}^{n} \alpha_i \phi(x_i).$$

Why is this? (Hint: you should not need to do any long proofs to answer the question).

[2 marks]

- Using the result of question 3, derive the prediction $f^*(x)$ at some new point x using only kernels between the feature maps $k(x,x') = \langle \phi(x), \phi(x')\rangle_{\mathcal{H}}$. You will need the relations

$$\frac{\partial a^\top U a}{\partial a} = (U + U^\top)a, \qquad \frac{\partial v^\top a}{\partial a} = \frac{\partial a^\top v}{\partial a} = v.$$

[5 marks]

- When reviewing a paper that uses kernel ridge regression, you encounter the sentence "the reported performance is the best average validation set error obtained over a range of λ values". Explain why you are rejecting the paper, and give an alternative procedure for selecting λ and evaluating test performance.

[5 marks]

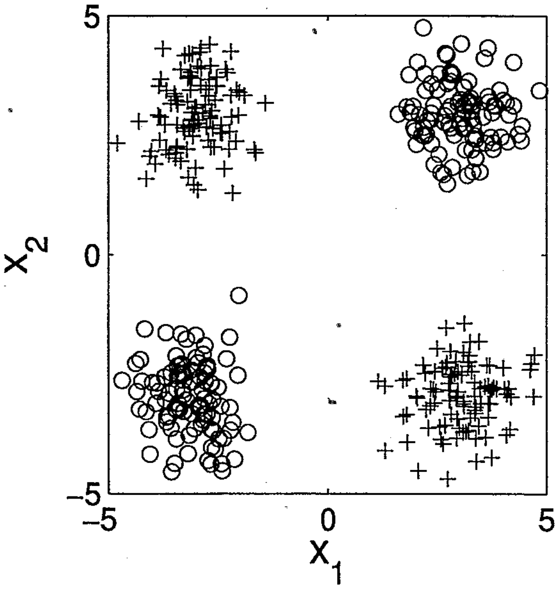


Figure 1: Figure for Question 2 part 1. The red crosses are from **P**, the blue circles are from **Q**.

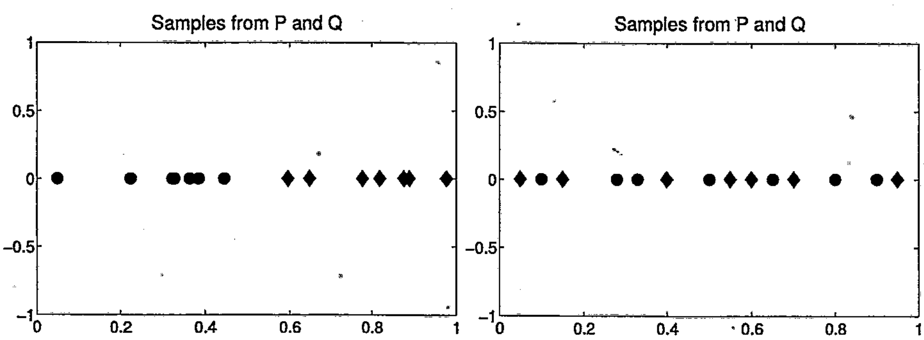


Figure 2: Figure for question 2 part 3. Samples from **P** are red circles, and from **Q** are blue diamonds.

2. • We are given independent and identically distributed samples $(x_i)_{i=1}^n$ from $\mathbf{P}$ and $(y_i)_{i=1}^n$ from $\mathbf{Q}$, defined on $\mathbb{R}^2$ as shown in the Figure 1. Write $x[1]$ the first coordinate of x, and $x[2]$ the second coordinate. Why wouldn't a t-test work in distinguishing $\mathbf{P}$ from $\mathbf{Q}$? Define a *finite dimensional* feature mapping, $\phi(x)$, such that $\mathbf{E}_\mathbf{P}(\phi(\mathrm{x})) \neq \mathbf{E}_\mathbf{Q}(\phi(\mathrm{y}))$, where x is the random variable with distribution $\mathbf{P}$ and y the random variable with distirbution $\mathbf{Q}$.

[2 marks]

• Define an RKHS $\mathcal{H}$ on $\mathcal{X}$ with feature map $\phi(x)$ kernel $k(x,x') = \langle\phi(x),\phi(x')\rangle_{\mathcal{H}}$. The Riesz representation theorem states that a bounded linear operator $A : \mathcal{H} \to \mathbb{R}$ has a representation $g_A \in \mathcal{H}$, such that $Af = \langle f, g_A\rangle_{\mathcal{H}}$. Show using this theorem that under a certain condition on k (what is it?), there exists a function $\mu_\mathbf{P} \in \mathcal{H}$ such that $\mathbf{E}_\mathbf{P} f(\mathrm{x}) = \langle \mu_\mathbf{P}, f\rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

[3 marks]

• A proposed measure of similarity of two probabilities is

$$\mathrm{MMD}\{\mathbf{P},\mathbf{Q};\mathcal{H}\} = \sup_{\|f\|_{\mathcal{H}}\leq 1} \left[\mathbf{E}_\mathbf{P} f(\mathrm{x}) - \mathbf{E}_\mathbf{Q} f(\mathrm{y})\right]. \tag{3}$$

Assume $\mathcal{H}$ is an RKHS with a Gaussian kernel. Which of the two distribution pairs in Figure 2 is further apart according to this measure, and why? Sketch (roughly) the f attaining the supremum in (3) for both cases.

[3 marks]

• Let $\mathcal{F}$ and $\mathcal{G}$ be reproducing kernel Hilbert spaces with respective kernels k and l, and feature maps ϕ and ψ. We are given an i.i.d. sample from $\mathbf{P}_{\mathrm{xy}}$, written $\mathbf{z} := ((x_1,y_1)\ldots(x_n,y_n))$. Derive a solution to

$$\mathrm{COCO} := \max_{f,g} \left\langle f, \widehat{C}_{XY} g\right\rangle_{\mathcal{F}}$$

$$\text{subject to} \quad \|f\|_{\mathcal{F}} = 1 \tag{4}$$

$$\|g\|_{\mathcal{G}} = 1, \tag{5}$$

where

$$\begin{aligned}\widehat{C}_{XY} &= \frac{1}{n}\sum_{i=1}^n (\phi(x_i) - \hat{\mu}_x) \otimes (\psi(y_i) - \hat{\mu}_y) \\ &= \frac{1}{n} X H Y^\top\end{aligned}$$

$H = I_n - n^{-1}\mathbf{1}_n$, for $\mathbf{1}_n$ an $n \times n$ matrix of ones,

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i) \qquad \hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n}\psi(y_i),$$

and

$$X = \left[\begin{array}{ccc} \phi(x_1) & \ldots & \phi(x_n) \end{array}\right] \qquad Y = \left[\begin{array}{ccc} \psi(y_1) & \ldots & \psi(y_n) \end{array}\right].$$

You may assume that

$$f = \sum_{i=1}^{n}\alpha_i\left[\phi(x_i) - \hat{\mu}_x\right] = XH\alpha \qquad g = \sum_{j=1}^{n}\beta_i\left[\psi(y_i) - \hat{\mu}_y\right] = YH\beta.$$

You may also need that $H = HH$.

[6 marks]

- Why is COCO a useful measure of dependence, even when the variables are dependent but have no linear correlation? You can sketch an example if you wish, but a clear explanation is sufficient.

[3 marks]

- COCO is a measure of dependence based on $\widehat{C}_{XY}$. What might be a better $\widehat{C}_{XY}$-based measure of dependence than COCO, and why?

[3 marks]

3. Assume $\mathcal{H}$ is a reproducing kernel Hilbert space with a Gaussian kernel,

$$k(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{\sigma}\right) \tag{6}$$

We have a sample $(x_i)_{i=1}^n$ drawn independently and identically from some distribution P. The one-class SVM requires a solution to the following problem:

$$\begin{aligned}\underset{w\in\mathcal{H},\xi\in\mathbb{R}^n,\rho\in\mathbb{R}}{\text{minimize}} \; & \frac{1}{2}\|w\|_{\mathcal{H}}^2 + \frac{1}{\nu n}\sum_{i=1}^n \xi_i - \rho \\ \text{subject to } & \langle w, \phi(x_i)\rangle_{\mathcal{H}} \geq \rho - \xi_i,\ \xi_i \geq 0,\end{aligned} \tag{7}$$

where $\nu \in (0,1]$ is a parameter.

- What does the function

$$f(x) = \text{sign}\left(\langle w, \phi(x)\rangle_{\mathcal{H}} - \rho\right)$$

  do (recall that the function *sign* returns $+1$ for positive arguments, -1 for negative arguments, and 0 when the argument is 0)? Hint: sketch the hyperplane returned by the optimization (7) in feature space. It will help to recall that in the Gaussian RKHS with kernel (6), all points are on the surface of the unit sphere, and the angle between any two points (measured wrt the origin) is no greater than $\pi/2$.

  [6 marks]

- Write the dual solution of this optimization problem. State the KKT conditions as they apply to this problem.

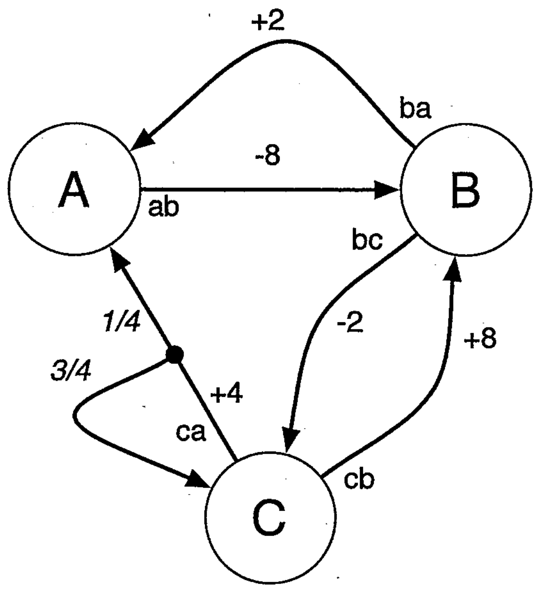
  [8 marks]

- Show that ν is an upper bound on the fraction of "outliers" (for which $\langle w, \phi(x_i)\rangle_{\mathcal{H}} < \rho$) lower bound on the fraction of SVs (points for which $\langle w, \phi(x_i)\rangle_{\mathcal{H}} \leq \rho$). Hint: if $w = \sum_{i=1}^n \alpha_i \phi(x_i)$, look at what the KKT conditions imply about α_i: in particular, how they determine the lower and upper bound on each α_i.

  [6 marks]

Part B: Reinforcement Learning

4. Consider the following Markov Decision Process (MDP) with discount factor $\gamma = 0.5$. Upper case letters A, B, C represent states; arcs represent state transitions; lower case letters ab, ba, bc, ca, cb represent actions; signed integers represent rewards; and fractions represent transition probabilities.



- Define the *state-value function* $V^{\pi}(s)$ for a discounted MDP

[1 marks]

- Write down the *Bellman expectation equation* for state-value functions

[2 marks]

- Consider the uniform random policy π_1(s,a) that takes all actions from state s with equal probability. Starting with an initial value function of $V_1(A) = V_1(B) = V_1(C) = 2$, apply one synchronous iteration of iterative policy evaluation (i.e. one backup for each state) to compute a new value function $V_2(s)$

[3 marks]

- Apply one iteration of greedy policy improvement to compute a new, deterministic policy $\pi_2(s)$

[2 marks]

- Consider a deterministic policy π(s). Prove that if a new policy π' is greedy with respect to V^{π} then it must be better than or equal to π, i.e. $V^{\pi'}(s) \geq V^{\pi}(s)$ for all s; and that if $V^{\pi'}(s) = V^{\pi}(s)$ for all s then π' must be an optimal policy.

[5 marks]

- Define the *optimal state-value function* $V^*(s)$ for an MDP

[1 marks]

- Write down the *Bellman optimality equation* for state-value functions

[2 marks]

- Starting with an initial value function of $V_1(A) = V_1(B) = V_1(C) = 2$, apply one synchronous iteration of value iteration (i.e. one backup for each state) to compute a new value function $V_2(s)$.

[3 marks]

- Is your new value function $V_2(s)$ optimal? Justify your answer.

[1 marks]

[Total 20 marks]

5. Consider an undiscounted Markov Reward Process with two states A and B. The transition matrix and reward function are unknown, but you have observed two sample episodes:

$$A+3 \rightarrow A+2 \rightarrow B-4 \rightarrow A+4 \rightarrow B-3 \rightarrow \text{terminate}$$

$$B-2 \rightarrow A+3 \rightarrow B-3 \rightarrow \text{terminate}$$

In the above episodes, sample state transitions and sample rewards are shown at each step, e.g. $A+3 \rightarrow A$ indicates a transition from state A to state A, with a reward of $+3$.

- Using first-visit Monte-Carlo evaluation, estimate the state-value function $V(A), V(B)$

[2 marks]

- Using every-visit Monte-Carlo evaluation, estimate the state-value function $V(A), V(B)$.

[2 marks]

- Draw a diagram of the Markov Reward Process that best explains these two episodes (i.e. the reward function/transition matrix corresponding to the empirical mean reward/transitions in the data). Show rewards and transition probabilities on your diagram.

[4 marks]

- Define the Bellman equation for a Markov reward process

[2 marks]

- Solve the Bellman equation to give the true state-value function $V(A), V(B)$. Hint: solve the Bellman equations directly, rather than iteratively.

[4 marks]

- What value function would batch TD(0) find, i.e. if TD(0) was applied repeatedly to these two episodes?

[2 marks]

- What value function would batch TD(1) find, using accumulating eligibility traces?

[2 marks]

- What value function would LSTD(0) find?

[2 marks]

[Total 20 marks]

6. A rat is involved in an experiment. It experiences one episode. At the first step it hears a bell. At the second step it sees a light. At the third step it both hears a bell and sees a light. It then receives some food, worth +1 reward, and the episode terminates on the fourth step. All other rewards were zero. The experiment is undiscounted.

- Represent the rat's state s by two binary features, $bell(s) \in \{0,1\}$ and $light(s) \in \{0,1\}$. Write down the sequence of features corresponding to this episode.

  [3 marks]

- Approximate the state-value function by a linear combination of these features with two parameters: $b \cdot bell(s) + l \cdot light(s)$. If $b = 2$ and $l = -2$ then write down the sequence of approximate values corresponding to this episode.

  [3 marks]

- Define the λ-return v_t^λ

  [1 marks]

- Write down the sequence of λ-returns v_t^λ corresponding to this episode, for $\lambda = 0.5$ and $b = 2, l = -2$ (Hint: ensure the weights on your n-step returns add up to 1 by weighting your final step appropriately)

  [3 marks]

- Using the forward-view TD(λ) algorithm and your linear function approximator, what are the sequence of updates to weight b? What is the total update to weight b? Use $\lambda = 0.5, \gamma = 1, \alpha = 0.5$ and start with $b = 2, l = -2$

  [3 marks]

- Define the TD(λ) *accumulating eligibility trace* $\mathbf{e_t}$ (1) when using linear value function approximation

  [1 marks]

- Write down the sequence of eligibility traces $\mathbf{e_t}$ corresponding to the bell, using $\lambda = 0.5, \gamma = 1$

  [3 marks]

- Using the backward-view TD(λ) algorithm and your linear function approximator, what are the sequence of updates to weight b? (Use offline updates, i.e. do not actually change your weights, just accumulate your updates). What is the total update to weight b? Use $\lambda = 0.5, \gamma = 1, \alpha = 0.5$ and start with $b = 2, l = -2$

[3 marks]

[Total 20 marks]

END OF PAPER