Indian Institute of Technology Kanpur
CS771 Introduction to Machine Learning

Instructor: Purushottam Kar
Date: October 31, 2017
Total: 120 marks

ASSIGNMENT 3

Instructions

1. Only electronic submissions to this assignment will be accepted. No handwritten, scanned or photographed submissions would be accepted. There is only one part to this assignment
2. Submissions will be accepted till November 14, 2017, 2359 hrs, IST.
3. Late submissions will be accepted till November 19, 2017, 2359 hrs, IST.
4. Late submissions will incur a penalty – the maximum marks for a late submission will be 80% of the total marks, even if solutions to all questions are absolutely correct.
5. We will be closely checking your submissions for instances of plagiarism.
6. You may be penalized if you do not follow the formatting instructions given below very carefully.

Theory Part Submission

1. Your submission should be a single PDF file. No zip/tar/png/jpg files will be accepted.
2. The PDF file should have been compiled using the LaTeX style file provided to you.
3. Your answer to every question should begin on a new page. The style file is designed to do this automatically. However, if it fails to do so, use the `\clearpage` option in LaTeX before starting the new question, to enforce this.
4. Submissions for this part should be made on Gradescope `https://gradescope.com`.
5. Late submissions for this part should be made on Gradescope itself.
6. An account has been created for you on this website. Use your IITK CC ID (not GMail, CSE etc IDs) to login and use the "Forgot Password" option to set your password initially.
7. While submitting your assignment on this website, you will have to specify on which page(s) is question 1 answered, on which page(s) is question 2 answered etc. To do this properly, first ensure that the answer to each question starts on a different page.
8. Be careful to flush all your floats (figures, tables) corresponding to question n before starting the answer to question $n+1$ otherwise graders might miss your figures and award you less points. Again, the style file should do this automatically but be careful.
9. Your solutions must appear in proper order in the PDF file i.e. your solution to question n must be complete in the PDF file (including all plots, tables, proofs etc) before you present a solution to question $n+1$.
10. We may impose a penalty on submissions that significantly deviate from the style file or which do not following the formatting instructions.

Problem 3.1 (A Consistency Crisis for EM!)**.** Refer to lecture 16 material for this exercise. Let us be given data $X = \left[\mathbf{x}^1, \ldots, \mathbf{x}^n\right]$ which redacts the identities of latent variables $\mathbf{z}^1, \ldots, \mathbf{z}^n$, with the task being to estimate the MLE model $\boldsymbol{\theta}^{\text{MLE}} \in \Theta$ such that $\boldsymbol{\theta}^{\text{MLE}} \in \arg\max_{\boldsymbol{\theta} \in \Theta} \mathbb{P}\left[X \,|\, \boldsymbol{\theta}\right]$. We have seen how, the EM algorithm proceeds by first finding an estimate $\boldsymbol{\theta}^t$, then constructing a "Q-function" $Q_{\boldsymbol{\theta}^t} : \Theta \to \mathbb{R}$ as

$$Q_{\boldsymbol{\theta}^t}(\boldsymbol{\theta}) = \sum_{i=1}^{n} Q_{i,\boldsymbol{\theta}^t}(\boldsymbol{\theta}),$$

where $Q_{i,\boldsymbol{\theta}^t}(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z} \sim \mathbb{P}\left[\mathbf{z} \,|\, \mathbf{x}^i, \boldsymbol{\theta}^t\right]} \log \mathbb{P}\left[\mathbf{x}^i, \mathbf{z} \,|\, \theta\right]$, and then updating $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta} \in \Theta} Q_{\boldsymbol{\theta}^t}(\boldsymbol{\theta})$. Given this, show the following *self-consistency* properties of the Q-function

1. $\boldsymbol{\theta}^{\text{MLE}} \in \arg\max_{\boldsymbol{\theta} \in \Theta} Q_{\boldsymbol{\theta}^{\text{MLE}}}(\boldsymbol{\theta})$
2. If $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2 \in \arg\max_{\boldsymbol{\theta} \in \Theta} \mathbb{P}\left[X \,|\, \boldsymbol{\theta}\right]$ are two distinct but optimal MLE solutions then $\boldsymbol{\theta}^1 \in \arg\max_{\boldsymbol{\theta} \in \Theta} Q_{\boldsymbol{\theta}^2}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^2 \in \arg\max_{\boldsymbol{\theta} \in \Theta} Q_{\boldsymbol{\theta}^1}(\boldsymbol{\theta})$

You may reuse results proved in class without proving them again but you must clearly point to the lecture number, slide number in which that result was proved and also state the result clearly before using it in your proofs. (7+8=15 marks)

Problem 3.2 (ReLU guys! I'm going home!)**.** In this exercise, we will show that a ReLU network always learns a piecewise linear function. An n-partition of a set $\mathcal{X}$ is a collection of n subsets $\{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ such that each $\mathcal{X}_i \subseteq \mathcal{X}$ and

- $\mathcal{X}_i \cap \mathcal{X}_j = \phi$ if $i \neq j$
- $\bigcup_{i=1}^{n} \mathcal{X}_i = \mathcal{X}$

A piecewise linear function $f : \mathbb{R}^d \to \mathbb{R}$ with $n > 0$ "pieces" is indexed by an n-partition $\{\Omega_1, \ldots, \Omega_n\}$ of $\mathbb{R}^d$ and n linear models $\mathbf{w}^1, \ldots, \mathbf{w}^n$ such that for any $\mathbf{x} \in \mathbb{R}^d$, we have

$$f(\mathbf{x}) = \sum_{i=1}^{n} \mathbb{I}\left\{\mathbf{x} \in \Omega_i\right\} \cdot \left\langle \mathbf{w}^i, \mathbf{x} \right\rangle,$$

where $\mathbb{I}\{E\} = 1$ if E is true and 0 otherwise. Now, let $f_{\text{ReLU}}(v) = \max(v, 0)$ for any $v \in \mathbb{R}$ denote the ReLU activation function. Then show that

1. For any piecewise linear function f, and any scalar $c \in \mathbb{R}$, the function $g(\mathbf{x}) = c \cdot f(\mathbf{x})$ is also piecewise linear.
2. The sum of two piecewise linear functions is piecewise linear. Be careful that the two functions could correspond to different (number of) partitions of $\mathbb{R}^d$.
3. For a piecewise linear function f, the function $g(\mathbf{x}) = f_{\text{ReLU}}(f(\mathbf{x}))$ is also piecewise linear.
4. Any neural network with a ReLU activation function computes a piecewise linear function.
5. **Bonus**: If the network d input nodes, only one hidden layer with D nodes and only one output node, and all nodes except input layer nodes apply the ReLU activation function, how many "pieces" does the function computed by the network correspond to?

(5+10+5+15 = 35 marks)

Problem 3.3 (Kernel Perceptron). Develop a variant of the perceptron algorithm that can work in an RKHS corresponding to a Mercer kernel K. Your algorithm is forbidden from explicitly computing the feature map corresponding to K even once. Your perceptron should at every time step (see lecture 10), receive a data point $(x^t, y^t) \in \mathcal{X} \times \{-1, +1\}$ and perform updates. Just state your final algorithm cleanly giving all details in pseudo-code format - no derivations needed. You are encouraged to use the `mlalgorithm` command to format your pseudo code neatly. See `https://piazza.com/class/j5toxxryhdx56k?cid=354` for help. (25 marks)

Problem 3.4 (A Kernel is All You Need). We will denote a 2-dimensional vector as $\mathbf{z} = (x, y) \in \mathbb{R}^2$ where $x, y \in \mathbb{R}$ are the coordinates of the point. Consider the quadratic kernel over these points

$$K(\mathbf{z}^1, \mathbf{z}^2) = (\langle \mathbf{z}^1, \mathbf{z}^2 \rangle + 1)^2.$$

Let $\mathcal{H}_K$ denote the RKHS of the kernel K and let φ_K be the feature map for K. A quadratic function over $\mathbb{R}^2$ is parameterized as $(A, \mathbf{b}, c) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^2 \times \mathbb{R}$ as

$$f_{(A,\mathbf{b},c)}(\mathbf{z}) = \langle \mathbf{z}, A\mathbf{z} \rangle + \langle \mathbf{b}, \mathbf{z} \rangle + c$$

1. Show that the kernel K is Mercer by giving an explicit construction for $\varphi : \mathbb{R}^2 \to \mathcal{H}_K$. You will need to set $\mathcal{H}_K \equiv \mathbb{R}^D$ for an appropriate value of D. What D did you choose?
2. For every quadratic function $f_{(A,\mathbf{b},c)}$ over $\mathbb{R}^2$, construct a $\mathbf{w} \in \mathcal{H}_K$ such that for all $\mathbf{z} \in \mathbb{R}^2$
$$f_{(A,\mathbf{b},c)}(\mathbf{z}) = \langle \mathbf{w}, \varphi_K(\mathbf{z}) \rangle$$
3. For every $\mathbf{w} \in \mathcal{H}_K$, construct a triplet $(A, \mathbf{b}, c) \in \mathbb{R}^{2\times 2} \times \mathbb{R}^2 \times \mathbb{R}$ such that for all $\mathbf{z} \in \mathbb{R}^2$
$$f_{(A,\mathbf{b},c)}(\mathbf{z}) = \langle \mathbf{w}, \varphi_K(\mathbf{z}) \rangle$$
4. Given a regression dataset $(Z, \mathbf{y}) \in \mathbb{R}^{2\times n} \times \mathbb{R}^n$, show that as the regularization parameter $\lambda \to 0^+$, the output of kernel ridge regression over $(Z, \mathbf{y})$ using the kernel K, is a quadratic function $\hat{f}$ over $\mathbb{R}^2$ that offers a least squares error that is arbitrarily close to the smallest least squares error achievable over the dataset by any quadratic function over $\mathbb{R}^2$ i.e.
$$\sum_{i=1}^{n} (y^i - \hat{f}(\mathbf{z}^i))^2 \leq \min_{(A,\mathbf{b},c)\in\mathbb{R}^{2\times 2}\times\mathbb{R}^2\times\mathbb{R}} \sum_{i=1}^{n} (y^i - f_{(A,\mathbf{b},c)}(\mathbf{z}^i))^2 + \epsilon$$

where $\epsilon \to 0$ as $\lambda \to 0^+$. (10+7+8+10=35 marks)

Problem 3.5 (Why PCA does Mean-centering). Recall that we advocated a mean-centering pre-processing step to ensure optimal performance for PCA and PPCA routines. Lets see why is it that the mean is chosen to center. Suppose the low-dimensional latent factors are generated as

$$\mathbb{P}[\mathbf{z}] = \mathcal{N}(\mathbf{0}, I_k) \in \mathbb{R}^k,$$

whereupon an affine transformation is applied to them and noise is added to produce the observed data point, i.e. for $W \in \mathbb{R}^{d\times k}, \boldsymbol{\mu} \in \mathbb{R}^d, \sigma \geq 0$

$$\mathbb{P}[\mathbf{x} \,|\, \mathbf{z}] = \mathcal{N}(\mathbf{x} \,|\, W\mathbf{z} + \mu, \sigma^2 \cdot I_d) \in \mathbb{R}^d.$$

Note that the transformation is affine $W\mathbf{z}^i + \mu$ instead of linear in this example. Now using conjugacy properties of the Gaussian (see [**BIS**] Chapter 12), we can show that

$$\mathbb{P}[\mathbf{x}] = \int_{\mathbf{z}} \mathbb{P}[\mathbf{x} \,|\, \mathbf{z}]\, \mathbb{P}[\mathbf{z}] \; d\mathbf{z} = \mathcal{N}(\mathbf{x} \,|\, \boldsymbol{\mu}, C),$$

where $C = WW^\top + \sigma^2 \cdot I_d$. For a dataset $X = \left[\mathbf{x}^1, \ldots, \mathbf{x}^n\right]$, write down the complete expression for the data log-likelihood $\mathbb{P}[X \,|\, \boldsymbol{\mu}, W, \sigma]$. Do not ignore constants in your expression. Then derive an expression for $\boldsymbol{\mu}^{\mathrm{MLE}} = \arg\max_{\boldsymbol{\mu}\in\mathbb{R}^d} \mathbb{P}[X \,|\, \boldsymbol{\mu}, W, \sigma]$. Show all steps. (3+7=10 marks)