The Australian National University
School of Computing
Liang Zheng

Semester 2, 2022
Theory Assignment 2 of 4

COMP3670/6670: Introduction to Machine Learning

Release Date. 17 August 2022
Due Date. 00:30am, 19 September 2022
Maximum credit. 100

Exercise 1 **Conjectures** 5 credits each

Here are a collection of conjectures. Which are true, and which are false?

- If it is true, provide a formal proof demonstrating so.
- If it is false, give a counterexample, clearly stating why your counterexamples satisfies the premise but not the conclusion.

(No marks for just starting True/False.)

Hint: There's quite a few questions here, but each is relatively simple (the counterexamples aren't very complicated, and the proofs are short.) Try playing around with a few examples first to get an intuitive feeling if the statement is true before trying to prove it.

Let V be a vector space, and let $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ be an inner product over V.

1. Triangle inequality for inner products: For all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in V$, $\langle \mathbf{a}, \mathbf{c} \rangle \leq \langle \mathbf{a}, \mathbf{b} \rangle + \langle \mathbf{b}, \mathbf{c} \rangle$.
2. Transitivity of orthogonality: For all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in V$, if $\langle \mathbf{a}, \mathbf{b} \rangle = 0$ and $\langle \mathbf{b}, \mathbf{c} \rangle = 0$ then $\langle \mathbf{a}, \mathbf{c} \rangle = 0$.
3. Orthogonality closed under addition: Suppose $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subseteq V$ is a set of vectors, and $\mathbf{x}$ is orthogonal to all of them (that is, for all $i = 1, 2, \ldots n$, $\langle \mathbf{x}, \mathbf{v_i} \rangle = 0$). Then $\mathbf{x}$ is orthogonal to any $\mathbf{y} \in \text{Span}(S)$.
4. Let $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \subseteq V$ be an **orthonormal** set of vectors in V. Then for all **non-zero** $\mathbf{x} \in V$, if for all $1 \leq i \leq n$ we have $\langle \mathbf{x}, \mathbf{v_i} \rangle = 0$ then $\mathbf{x} \notin \text{Span}(S)$.
5. Let $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \subseteq V$ be a set of vectors in V (no assumption of orthonormality). Then for all **non-zero** $\mathbf{x} \in V$, if for all $1 \leq i \leq n$ we have $\langle \mathbf{x}, \mathbf{v}_i \rangle = 0$ then $\mathbf{x} \notin \text{Span}(S)$.
6. Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a set of **orthonormal** vectors such that $\text{Span}(S) = V$, and let $\mathbf{x} \in V$. Then there is a *unique* set of coefficients $c_1, \ldots, c_n$ such that
$$\mathbf{x} = c_1\mathbf{v}_1 + \ldots + c_n\mathbf{v}_n$$
7. Let $S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ be a set of vectors (no assumption of orthonormality) such that $\text{Span}(S) = V$, and let $\mathbf{x} \in V$. Then there is a *unique* set of coefficients $c_1, \ldots, c_n$ such that
$$\mathbf{x} = c_1\mathbf{v}_1 + \ldots + c_n\mathbf{v}_n$$
8. Let $S = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \subseteq V$ be a set of vectors. If all the vectors are pairwise linearly independent (i.e., for any $1 \leq i \neq j \leq n$, then only solution to $c_i\mathbf{v}_i + c_j\mathbf{v}_j = \mathbf{0}$ is the trivial solution $c_i = c_j = 0$.) then the set S is linearly independent.

Exercise 2 **Inner Products induce Norms** 20 credits

Let V be a vector space, and let $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ be an inner product on V. Define $||\mathbf{x}|| := \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. Prove that $|| \cdot ||$ is a norm.

(Hint: To prove the triangle inequality holds, you may need the Cauchy-Schwartz inequality, $\langle \mathbf{x}, \mathbf{y} \rangle \leq ||\mathbf{x}|| ||\mathbf{y}||$.)

Exercise 3 **General Linear Regression with Regularisation** (10+10+10+5+5 credits)

Let $\mathbf{A} \in \mathbb{R}^{N \times N}, \mathbf{B} \in \mathbb{R}^{D \times D}$ be *symmetric, positive definite* matrices. From the lectures, we can use symmetric positive definite matrices to define a corresponding inner product, as shown below. We can also define a norm using the inner products.

$$\begin{aligned} \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{A}} &:= \mathbf{x}^T \mathbf{A} \mathbf{y} \\ \|\mathbf{x}\|_{\mathbf{A}}^2 &:= \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{A}} \\ \langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{B}} &:= \mathbf{x}^T \mathbf{B} \mathbf{y} \\ \|\mathbf{x}\|_{\mathbf{B}}^2 &:= \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{B}} \end{aligned}$$

Suppose we are performing linear regression, with a training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where for each i, $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$. We can define the matrix

$$\boldsymbol{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times D}$$

and the vector

$$\mathbf{y} = [y_1, \ldots, y_N]^T \in \mathbb{R}^N.$$

We would like to find $\boldsymbol{\theta} \in \mathbb{R}^D, \mathbf{c} \in \mathbb{R}^N$ such that $\mathbf{y} \approx \mathbf{X}\boldsymbol{\theta} + \mathbf{c}$, where the error is measured using $\| \cdot \|_{\mathbf{A}}$. We avoid overfitting by adding a weighted regularization term, measured using $||\cdot||_{\mathbf{B}}$. We define the loss function with regularizer:

$$\mathcal{L}_{\mathbf{A},\mathbf{B},\mathbf{y},\mathbf{X}}(\boldsymbol{\theta}, \mathbf{c}) = ||\mathbf{y} - \boldsymbol{X}\boldsymbol{\theta} - \mathbf{c}||_{\mathbf{A}}^2 + ||\boldsymbol{\theta}||_{\mathbf{B}}^2 + \|\mathbf{c}\|_{\mathbf{A}}^2$$

For the sake of brevity we write $\mathcal{L}(\boldsymbol{\theta}, \mathbf{c})$ for $\mathcal{L}_{\mathbf{A},\mathbf{B},\mathbf{y},\mathbf{X}}(\boldsymbol{\theta}, \mathbf{c})$.

HINTS:

- You may use (without proof) the property that a symmetric positive definite matrix is invertible.
- We assume that there are sufficiently many non-redundant data points for $\mathbf{X}$ to be full rank. In particular, you may assume that the null space of $\mathbf{X}$ is trivial (that is, the only solution to $\mathbf{X}\mathbf{z} = \mathbf{0}$ is the trivial solution, $\mathbf{z} = \mathbf{0}$.)
- You may use identities of gradients from the lectures slides, so long as you mention as such.

1. Find the gradient $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{c})$.
2. Let $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{c}) = \mathbf{0}$, and solve for $\boldsymbol{\theta}$. If you need to invert a matrix to solve for $\boldsymbol{\theta}$, you should prove the inverse exists.
3. Find the gradient $\nabla_{\mathbf{c}} \mathcal{L}(\boldsymbol{\theta}, \mathbf{c})$.

   We now compute the gradient with respect to $\mathbf{c}$.
4. Let $\nabla_{\mathbf{c}} \mathcal{L}(\boldsymbol{\theta}) = \mathbf{0}$, and solve for $\mathbf{c}$. If you need to invert a matrix to solve for $\mathbf{c}$, you should prove the inverse exists.
5. Show that if we set $\mathbf{A} = \mathbf{I}, \mathbf{c} = \mathbf{0}, \mathbf{B} = \lambda \mathbf{I}$, where $\lambda \in \mathbb{R}$, your answer for 3.2 agrees with the analytic solution for the standard least squares regression problem with L2 regularization, given by

$$\boldsymbol{\theta} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \mathbf{y}.$$