Homework 6 Solutions

Math 198: Math for Machine Learning

Due Date:
Name:
Student ID:

1 Ridge Regression and Kernel Trick

1. (Adapted from CS189 Fa19 HW2.) Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a data matrix, $\mathbf{y} \in \mathbb{R}^n$ be an observation vector, and $\mathbf{w}_\lambda \in \mathbb{R}^d$ be the ridge regression solution, i.e., $\mathbf{w}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$. Furthermore, let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \sum\limits_{i=1}^{d} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$ be the SVD of $\mathbf{X}$.

   (a) Show that $\mathbf{w}_\lambda = \sum\limits_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2+\lambda} \mathbf{v}_i \langle \mathbf{u}_i, \mathbf{y} \rangle$.

$$
\begin{aligned}
\mathbf{w}_\lambda &= (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \\
&= ((\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top + \lambda \mathbf{I})^{-1} (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top \mathbf{y} \\
&= (\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top + \lambda \mathbf{V}\mathbf{V}^\top)^{-1} \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top \mathbf{y} \text{ by unitarity of } \mathbf{V} \\
&= (\sum_{i=1}^{d} \sigma_i^2 \mathbf{v}_i \mathbf{v}_i^\top + \lambda \mathbf{v}_i \mathbf{v}_i^\top)^{-1} \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top \mathbf{y} \\
&= (\sum_{i=1}^{d} (\sigma_i^2 + \lambda) \mathbf{v}_i \mathbf{v}_i^\top)^{-1} \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top \mathbf{y} \\
&= (\mathbf{V}(\mathbf{\Sigma}^2 + \lambda \mathbf{I})\mathbf{V}^\top)^{-1} \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top \mathbf{y} \\
&= \mathbf{V}(\mathbf{\Sigma}^2 + \lambda \mathbf{I})^{-1}\mathbf{V}^\top \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top \mathbf{y} \\
&= \mathbf{V}(\mathbf{\Sigma}^2 + \lambda \mathbf{I})^{-1}\mathbf{\Sigma}\mathbf{U}^\top \mathbf{y}
\end{aligned}
$$

   Observe that $(\mathbf{\Sigma}^2 + \lambda \mathbf{I})^{-1}\mathbf{\Sigma}$ is a diagonal matrix with entries $\frac{\sigma_i}{\sigma_i^2+\lambda}$ on the diagonal. Therefore, we can write

$$
\begin{aligned}
\mathbf{w}_\lambda &= \mathbf{V}(\mathbf{\Sigma}^2 + \lambda \mathbf{I})^{-1}\mathbf{\Sigma}\mathbf{U}^\top \mathbf{y} \\
&= \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2 + \lambda} \mathbf{v}_i \langle \mathbf{u}_i, \mathbf{y} \rangle
\end{aligned}
$$

   completing the proof.

   (b) Deduce that the OLS solution $\mathbf{w}_{\text{OLS}} = \sum\limits_{i=1}^{d} \frac{1}{\sigma_i} \mathbf{v}_i \langle \mathbf{u}_i, \mathbf{y} \rangle$.

   The OLS solution is identical to the ridge regression solution, except with $\lambda = 0$. Setting λ to 0 in the ridge regression solution from (a) yields this solution.

   (c) Prove that $\lim\limits_{\lambda \to 0} \mathbf{w}_\lambda = \mathbf{w}_{\text{OLS}}$.

$$\lim_{\lambda \to 0} \mathbf{w}_\lambda = \lim_{\lambda \to 0} \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2 + \lambda} \mathbf{v}_i \langle \mathbf{u}_i, \mathbf{y} \rangle$$
$$= \sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2} \mathbf{v}_i \langle \mathbf{u}_i, \mathbf{y} \rangle$$
$$= \mathbf{w}_{\text{OLS}}$$

(d) Show that if $\mathbf{w}_\lambda \neq 0$, then the map $\lambda \to ||\mathbf{w}_\lambda||^2$ is strictly decreasing and strictly positive on $(0, \infty)$. What is the effect of λ on $\mathbf{w}_\lambda$?

$$\begin{aligned}
||\mathbf{w}_\lambda||^2 &= ||\sum_{i=1}^{d} \frac{\sigma_i}{\sigma_i^2 + \lambda} \mathbf{v}_i \langle \mathbf{u}_i, \mathbf{y} \rangle||^2 \\
&= \sum_{i=1}^{d} ||\frac{\sigma_i}{\sigma_i^2 + \lambda} \mathbf{v}_i \langle \mathbf{u}_i, \mathbf{y} \rangle||^2 + \sum_{i \neq j} \frac{\sigma_i}{\sigma_i^2 + \lambda} \frac{\sigma_j}{\sigma_j^2 + \lambda} \mathbf{v}_i^\top \mathbf{v}_j \langle \mathbf{u}_i, \mathbf{y} \rangle \langle \mathbf{u}_j, \mathbf{y} \rangle \\
&= \sum_{i=1}^{d} ||\frac{\sigma_i}{\sigma_i^2 + \lambda} \mathbf{v}_i \langle \mathbf{u}_i, \mathbf{y} \rangle||^2 \text{ since } \mathbf{v}_i^\top \mathbf{v}_j = 0 \text{ for } i \neq j \\
&= \sum_{i=1}^{d} (\frac{\sigma_i}{\sigma_i^2 + \lambda} \langle \mathbf{u}_i, \mathbf{y} \rangle)^2 \text{ since } \mathbf{v}_i^\top \mathbf{v}_i = 1
\end{aligned}$$

Because this is a sum of squares, it is always positive; furthermore, as λ increases, the denominator of each term increases, and thus $||\mathbf{w}_\lambda||^2$ decreases. Therefore, λ reduces the norm of the solution, and so higher values of λ will produce less complex weights.

2. Prove that the kernel trick holds for cubic polynomials in two variables. That is, if the feature map ϕ maps

$$\begin{bmatrix} a_i & b_i \end{bmatrix}^\top \mapsto \begin{bmatrix} a_i^3 & b_i^3 & \sqrt{3}a_i^2 b_i & \sqrt{3}a_i b_i^2 & \sqrt{3}a_i^2 & \sqrt{3}b_i^2 & \sqrt{6}a_i b_i & \sqrt{3}a_i & \sqrt{3}b_i & 1 \end{bmatrix}^\top$$

then $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^\top \mathbf{x}_j + 1)^3$.

$$\begin{aligned}
k(\mathbf{x}_i, \mathbf{x}_j) &= \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) \\
&= a_i^3 a_j^3 + b_i^3 b_j^3 + 3a_i^2 b_i a_j^2 b_j + 3a_i b_i^2 a_j b_j^2 + 3a_i^2 a_j^2 + 3b_i^2 b_j^2 + 6a_i b_i a_j b_j + 3a_i a_j + 3b_i b_j + 1 \\
&= (a_i^3 a_j^3 + 3a_i^2 a_j^2 b_i b_j + 3a_i a_j b_i^2 b_j^2 + b_i^3 b_j^3) + 3(a_i^2 a_j^2 + 2a_i a_j b_i b_j + b_i^2 b_j^2) + 3(a_i a_j + b_i b_j) + 1 \\
&= (\mathbf{x}_i^\top \mathbf{x}_j)^3 + 3(\mathbf{x}_i^\top \mathbf{x}_j)^2 + 3(\mathbf{x}_i^\top \mathbf{x}_j) + 1 \\
&= (\mathbf{x}_i^\top \mathbf{x}_j + 1)^3
\end{aligned}$$

2 Linear Algebra Review

1. Let V be an arbitrary vector space. Prove that the zero vector $\mathbf{0} \in V$ is unique. Additionally, prove that for any vector $\mathbf{v} \in V$, the additive inverse $-\mathbf{v}$ is unique.
Suppose towards a contradiction that $\mathbf{0}$ is not unique. Then a second, distinct zero vector $\mathbf{0}' \in V$, exists. But then $\mathbf{0} = \mathbf{0} + \mathbf{0}' = \mathbf{0}'$. So $\mathbf{0}'$ is not distinct as claimed; therefore $\mathbf{0}$ is unique.
Suppose towards a contradiction that the additive inverse of $\mathbf{v}$ is not unique. Then a second, distinct vector $\mathbf{w} \in V$ exists such that $\mathbf{v} + \mathbf{w} = \mathbf{0}$. But then $\mathbf{w} = \mathbf{0} - \mathbf{v} = -\mathbf{v}$. So $\mathbf{w}$ is not distinct as claimed; therefore $-\mathbf{v}$ is unique up to $\mathbf{v}$.

2. Prove that the dot product is a valid inner product on $\mathbb{R}^n$.
Let $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$.

(a) Linearity (first coordinate). $(a\mathbf{u}+\mathbf{v})\cdot\mathbf{w} = \sum_{i=1}^{n}(au_i+v_i)w_i = \sum_{i=1}^{n} au_iw_i + v_iw_i = (a\mathbf{u}\cdot\mathbf{w}) + (\mathbf{v}\cdot\mathbf{w})$

(b) Symmetry. $\mathbf{v}\cdot\mathbf{w} = \sum_{i=1}^{n} v_iw_i = \sum_{i=1}^{n} w_iv_i = \mathbf{w}\cdot\mathbf{v}$

(c) PSD. $\mathbf{v}\cdot\mathbf{v} = \sum_{i=1}^{n} v_i^2 \geq 0$; $\mathbf{v}\cdot\mathbf{v} = 0 \iff \sum_{i=1}^{n} v_i^2 = 0 \iff \forall v_i, v_i = 0 \iff \mathbf{v} = \mathbf{0}$.

3. Let V and W be arbitrary vector spaces. Prove that $\dim V = \dim W$ if and only if there exists an isomorphism $f : V \to W$.
($\Rightarrow$) Let $\beta = \{\ldots, \beta_i, \ldots\}$ be a basis for V and $\gamma = \{\ldots, \gamma_i, \ldots\}$ a basis for W. Define $f : V \to W$ such that $\beta_i \mapsto \gamma_i$. This definition is valid, as there are as many elements in β as in γ (since the dimensions of V and W are equal). Furthermore, we can define $f^{-1} : W \to V$ such that $\gamma_i \mapsto \beta_i$. Then $\forall \mathbf{v} \in V, f^{-1}(f(\mathbf{v})) = \mathbf{v}$ (since $\mathbf{v}$ can be written as a sum of the basis elements in β, and the action of $f^{-1}(f(\cdot))$ is to switch the β_i's in that sum to γ_i's and back). So an isomorphism f exists.
($\Leftarrow$) Because f is onto, any vector $\mathbf{w} \in W$ can be written as $\mathbf{w} = f(\mathbf{v})$ for some $\mathbf{v} \in V$. But $\mathbf{v} = \sum_i a_i\beta_i$, so $\mathbf{w} = f(\mathbf{v}) = \sum_i a_i f(\beta_i)$. So the set $f(\beta) = \{\ldots, f(\beta_i), \ldots\}$ spans W. Suppose towards a contradiction that this set is not linearly independent, i.e. $\sum_i b_i f(\beta_i) = 0$ for appropriate b_i. Let $\mathbf{w}_1 = \sum_i (b_i+1)f(\beta_i), \mathbf{w}_2 = \sum_i f(\beta_i)$. Then $\mathbf{w}_1 - \mathbf{w}_2 = 0$, and so $\mathbf{w}_1 = \mathbf{w}_2$. But f is an isomorphism, so $\sum_i b_i\beta_i = f^{-1}(\mathbf{w}_1) = f^{-1}(\mathbf{w}_2) = \sum_i \beta_i$. Because β is a basis set for V, this is impossible; therefore, $f(\beta)$ is linearly independent, and is thus a basis set for W. Because this set contains the same number of elements as β, the vector spaces they generate, V and W, have equal dimension.

4. Prove that trace is a linear map, i.e. $\mathrm{tr}(c\mathbf{A}+\mathbf{B}) = c\mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$.
$\mathrm{tr}(c\mathbf{A}+\mathbf{B}) = \sum_i (c\mathbf{A}+\mathbf{B})_{ii} = \sum_i c\mathbf{A}_i + \mathbf{B}_i = c\sum_i \mathbf{A}_i + \sum_j \mathbf{B}_j = c\mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$

5. Let $\mathbf{A}$ be a square matrix and λ an eigenvalue of $\mathbf{A}$. Prove that λ^k is an eigenvalue of $\mathbf{A}^k$.
Suppose $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ corresponding to eigenvalue λ. Then $\mathbf{A}^k\mathbf{v} = \mathbf{A}\ldots\mathbf{A}\mathbf{v} = \mathbf{A}\ldots\lambda\mathbf{v} = \lambda^k\mathbf{v}$. So $\mathbf{v}$ is also an eigenvector of $\mathbf{A}^k$ corresponding to eigenvalue λ^k.

6. (Adapted from CS189 Fa19 HW0.) Let $\mathbf{v}$ and $\mathbf{w}$ be vectors in $\mathbb{R}^n$. Define $\mathbf{A} = \mathbf{v}\mathbf{w}^\top$. Find the non-zero eigenvalues of $\mathbf{A}$ and their eigenvectors, and determine the rank of the nullspace of $\mathbf{A}$.
We have

$$\mathbf{A} = \begin{bmatrix} \ldots & | & \ldots \\ \ldots & w_i\mathbf{v} & \ldots \\ \ldots & | & \ldots \end{bmatrix}$$

Clearly, the columns of $\mathbf{A}$ are not linearly independent, so $\mathbf{A}$ is not full rank. Since they are all spanned by only one vector, $\mathrm{rank}(\mathbf{A}) = 1$, so $\dim\ker(\mathbf{A}) = n-1$. This implies that there is only one non-zero eigenvector; by observation, $\mathbf{v}$ is an eigenvector, since $\mathbf{A}\mathbf{v} = \mathbf{v}\mathbf{w}^\top\mathbf{v} = \langle\mathbf{w},\mathbf{v}\rangle\mathbf{v}$; the corresponding eigenvalue is $\langle\mathbf{w},\mathbf{v}\rangle$.

7. Prove that a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is PSD if and only if there exists a matrix $\mathbf{U} \in \mathbb{R}^{n\times n}$ such that $\mathbf{A} = \mathbf{U}\mathbf{U}^\top$.
($\Rightarrow$) Because $\mathbf{A}$ is PSD, we can take the spectral decomposition $A = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ for unitary $\mathbf{U}$, diagonal $\mathbf{\Lambda}$. Furthermore, $\mathbf{\Lambda}$ contains the eigenvalues of $\mathbf{A}$ on its diagonal. Because $\mathbf{A}$ is PSD, all these eigenvalues are non-negative, so the matrix $\mathbf{\Lambda}^{\frac{1}{2}}$ containing their square roots on its diagonal exists. But then $\mathbf{A} = \mathbf{U}\mathbf{\Gamma}^{\frac{1}{2}}\mathbf{\Gamma}^{\frac{1}{2}}\mathbf{U}^\top$, and since $\mathbf{\Gamma}^{\frac{1}{2}}$ is symmetric, then if $\mathbf{U}' = \mathbf{U}\mathbf{\Gamma}^{\frac{1}{2}}$, $\mathbf{A} = \mathbf{U}'\mathbf{U}'^\top$.
($\Leftarrow$) Let $\mathbf{v}$ be some vector. Then $\mathbf{v}^\top\mathbf{A}\mathbf{v} = \mathbf{v}^\top\mathbf{U}\mathbf{U}^\top\mathbf{v} = \langle\mathbf{U}^\top\mathbf{v}, \mathbf{U}^\top\mathbf{v}\rangle = ||\mathbf{U}^\top\mathbf{v}||^2 \geq 0$. So $\mathbf{A}$ is PSD.