Homework 7

Deadline: Wednesday, Dec. 5, at 11:59pm.

Submission: You need to submit your solutions through MarkUs[1] as the PDF file `hw7_writeup.pdf`.

Neatness Point: One of the 10 points will be given for neatness. You will receive this point as long as we don't have a hard time reading your solutions or understanding the structure of your code.

Late Submission: 10% of the marks will be deducted for each day late, up to a maximum of 3 days. After that, no submissions will be accepted.

Collaboration. Weekly homeworks are individual work. See the Course Information handout[2] for detailed policies.

1. **[5pts] Representer Theorem.** In this question, you'll prove and apply a simplified version of the Representer Theorem, which is the basis for a lot of kernelized algorithms. Consider a linear model:

$$\begin{aligned} z &= \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}) \\ y &= g(z), \end{aligned}$$

where $\boldsymbol{\psi}$ is a feature map and g is some function (e.g. identity, logistic, etc.). We are given a training set $\{(\mathbf{x}^{(i)}, t^{(i)})\}_{i=1}^N$. We are interested in minimizing the expected loss plus an L_2 regularization term:

$$\mathcal{J}(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}(y^{(i)}, t^{(i)}) \; + \; \frac{\lambda}{2}\|\mathbf{w}\|^2,$$

where $\mathcal{L}$ is some loss function. Let $\boldsymbol{\Psi}$ denote the feature matrix

$$\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\psi}(\mathbf{x}^{(1)})^\top \\ \vdots \\ \boldsymbol{\psi}(\mathbf{x}^{(N)})^\top \end{pmatrix}.$$

Observe that this formulation captures a lot of the models we've covered in this course, including linear regression, logistic regression, and SVMs.

(a) **[2pts]** Show that the optimal weights must lie in the row space of $\boldsymbol{\Psi}$.
Hint: Given a subspace $\mathcal{S}$, a vector $\mathbf{v}$ can be decomposed as $\mathbf{v} = \mathbf{v}_\mathcal{S} + \mathbf{v}_\perp$, where $\mathbf{v}_\mathcal{S}$ is the projection of $\mathbf{v}$ onto $\mathcal{S}$, and $\mathbf{v}_\perp$ is orthogonal to $\mathcal{S}$. (You may assume this fact without proof, but you can review it here[3].) Apply this decomposition to $\mathbf{w}$ and see if you can show something about one of the two components.

[1] `https://markus.teach.cs.toronto.edu/csc411-2018-09`
[2] `http://www.cs.toronto.edu/~rgrosse/courses/csc411_f18/syllabus.pdf`
[3] `https://metacademy.org/graphs/concepts/projection_onto_a_subspace`

(b) **[3pts]** Another way of stating the result from part (a) is that $\mathbf{w} = \mathbf{\Psi}^\top \boldsymbol{\alpha}$ for some vector $\boldsymbol{\alpha}$. Hence, instead of solving for $\mathbf{w}$, we can solve for $\boldsymbol{\alpha}$. Consider the vectorized form of the L_2 regularized linear regression cost function:

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2N}\|\mathbf{t} - \mathbf{\Psi}\mathbf{w}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2.$$

Substitute in $\mathbf{w} = \mathbf{\Psi}^\top \boldsymbol{\alpha}$, to write the cost function as a function of $\boldsymbol{\alpha}$. Determine the optimal value of $\boldsymbol{\alpha}$. Your answer should be an expression involving λ, $\mathbf{t}$, and the Gram matrix $\mathbf{K} = \mathbf{\Psi}\mathbf{\Psi}^\top$. For simplicity, you may assume that $\mathbf{K}$ is positive definite. (The algorithm still works if $\mathbf{K}$ is merely PSD, it's just a bit more work to derive.)

Hint: the cost function $\mathcal{J}(\boldsymbol{\alpha})$ is a quadratic function. Simplify the formula into the following form:

$$\tfrac{1}{2}\boldsymbol{\alpha}^\top \mathbf{A}\boldsymbol{\alpha} + \mathbf{b}^\top \boldsymbol{\alpha} + c,$$

for some positive definite matrix $\mathbf{A}$, vector $\mathbf{b}$ and constant c (which can be ignored). You may assume without proof that the minimum of such a quadratic function is given by $\boldsymbol{\alpha} = -\mathbf{A}^{-1}\mathbf{b}$.

2. **[4pts] Compositional Kernels.** One of the most useful facts about kernels is that they can be composed using addition and multiplication. I.e., the sum of two kernels is a kernel, and the product of two kernels is a kernel. We'll show this in the case of kernels which represent dot products between finite feature vectors.

(a) **[1pt]** Suppose $k_1(x, x') = \boldsymbol{\psi}_1(x)^\top \boldsymbol{\psi}_1(x')$ and $k_2(x, x') = \boldsymbol{\psi}_2(x)^\top \boldsymbol{\psi}_2(x')$. Let k_S be the sum kernel $k_\mathrm{S}(x, x') = k_1(x, x') + k_2(x, x')$. Find a feature map $\boldsymbol{\psi}_\mathrm{S}$ such that $k_\mathrm{S}(x, x') = \boldsymbol{\psi}_\mathrm{S}(x)^\top \boldsymbol{\psi}_\mathrm{S}(x')$.

(b) **[3pts]** Suppose $k_1(x, x') = \boldsymbol{\psi}_1(x)^\top \boldsymbol{\psi}_1(x')$ and $k_2(x, x') = \boldsymbol{\psi}_2(x)^\top \boldsymbol{\psi}_2(x')$. Let k_P be the product kernel $k_\mathrm{P}(x, x') = k_1(x, x')\, k_2(x, x')$. Find a feature map $\boldsymbol{\psi}_\mathrm{P}$ such that $k_\mathrm{P}(x, x') = \boldsymbol{\psi}_\mathrm{P}(x)^\top \boldsymbol{\psi}_\mathrm{P}(x')$.

Hint: For inspiration, consider the quadratic kernel from Lecture 20, Slide 11.