Math Foundations of ML, Fall 2018

Homework #4

Due Monday September 24, at the beginning of class

As stated in the syllabus, unauthorized use of previous semester course materials is strictly prohibited in this course.

1. Using you class notes, prepare a 1-2 paragraph summary of what we talked about in class in the last week. I do not want just a bulleted list of topics, I want you to use complete sentences and establish context (Why is what we have learned relevant? How does it connect with other things you have learned here or in other classes?). The more insight you give, the better.

2. (a) The vector space $L_2([0,1]^2)$ is the space of signals of two variables, $x(s,t)$ with $s,t \in [0,1]$ such that
$$\int_0^1 \int_0^1 |x(s,t)|^2 \, \mathrm{d}s \, \mathrm{d}t \; < \; \infty.$$
Let $\{\psi_k(t),\ k \geq 0\}$ be an orthobasis for $L_2([0,1])$. Define
$$v_{k,\ell}(s,t) = \psi_k(s)\,\psi_\ell(t), \quad k,\ell \geq 0.$$
Show that $\{v_{k,\ell}(s,t),\ k,\ell \geq 0\}$ is an orthobasis for $L_2([0,1]^2)$. You need to argue that the $\boldsymbol{v}_{k,\ell}$ are orthonormal and that they span $L_2([0,1]^2)$.

   (b) Given on orthobasis for $L_2([0,1])$, describe how to construct an orthobasis for $L_2([0,1]^D)$ — the space of functions of D continuous-valued variables $x(\boldsymbol{t})$ such that
$$\int_0^1 \cdots \int_0^1 |x(\boldsymbol{t})|^2 \, \mathrm{d}t_1 \cdots \mathrm{d}t_D \; < \; \infty.$$

3. Let $\{\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_N\}$ be a basis for an N dimensional Hilbert space $\mathcal{S}$ with inner product $\langle \cdot, \cdot \rangle_S$. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ be specified by
$$\boldsymbol{x} = \sum_{n=1}^N \alpha_n \boldsymbol{\psi}_n, \quad \boldsymbol{y} = \sum_{n=1}^N \beta_n \boldsymbol{\psi}_n.$$
Fill in the blank:
$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_S \; = \; \underline{\hspace{2cm}}$$
with an expression that depends only on $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and the Gram matrix $\boldsymbol{G}$ for the basis.

4. Consider the same space $\mathcal{S}$ as in the previous question, and let $\boldsymbol{x} \in \mathcal{S}$ be given by
$$\boldsymbol{x} = \sum_{n=1}^N \alpha_n \boldsymbol{\psi}_n.$$

Since $\boldsymbol{x}$ is of course the closest point to itself in all of $\mathcal{S}$, we know that we can calculate the $\boldsymbol{\alpha}$ using

$$\boldsymbol{\alpha} = \boldsymbol{G}^{-1}\boldsymbol{b}, \quad \text{where} \quad \boldsymbol{b} = \begin{bmatrix} \langle \boldsymbol{x}, \boldsymbol{\psi}_1 \rangle_S \\ \langle \boldsymbol{x}, \boldsymbol{\psi}_2 \rangle_S \\ \vdots \\ \langle \boldsymbol{x}, \boldsymbol{\psi}_N \rangle_S. \end{bmatrix}$$

If we let $\boldsymbol{H} = \boldsymbol{G}^{-1}$, another way to write this is

$$\begin{aligned} \alpha_n &= \sum_{\ell=1}^{N} H_{n,\ell} \langle \boldsymbol{x}, \boldsymbol{\psi} \rangle_S \\ &= \left\langle \boldsymbol{x}, \sum_{\ell=1}^{N} H_{n,\ell} \boldsymbol{\psi}_\ell \right\rangle \\ &= \langle \boldsymbol{x}, \widetilde{\boldsymbol{\psi}}_n \rangle \end{aligned}$$

where

$$\widetilde{\boldsymbol{\psi}}_n = \sum_{\ell=1}^{N} H_{n,\ell} \boldsymbol{\psi}_\ell.$$

(a) Consider the basis from HW3:

$$\psi_n(t) = \phi\left(Nt - n + 1/2\right), \quad n = 1, \ldots, N, \quad \phi(t) = e^{-t^2},$$

for $N = 32$. Plot the dual basis vector $\widetilde{\psi}_{13}(t)$.

(b) If the elements of $\mathcal{S}$ are functions, we can also use the dual basis to write the reproducing kernel. We know that for any τ, and $\boldsymbol{s} \in \mathcal{S}$,

$$\begin{aligned} x(\tau) &= \sum_{n=1}^{N} \langle \boldsymbol{x}, \widetilde{\boldsymbol{\psi}}_n \rangle \psi_n(\tau) \\ &= \left\langle \boldsymbol{x}, \sum_{n=1}^{N} \psi_n(\tau)\, \widetilde{\boldsymbol{\psi}}_n \right\rangle \\ &= \langle \boldsymbol{x}, \boldsymbol{k}_\tau \rangle, \end{aligned}$$

where

$$\boldsymbol{k}_\tau = \sum_{n=1}^{N} \psi_n(\tau)\, \widetilde{\boldsymbol{\psi}}_n.$$

Plot $k_\tau(t)$ as a function of t for $\tau = .3242234$. Create a $\boldsymbol{x} \in \mathcal{S}$ by drawing the expansion coefficients $\boldsymbol{\alpha}$ at random (`alpha = randn(N,1);` in MATLAB), and verify that $\langle \boldsymbol{x}, \boldsymbol{k}_\tau \rangle = x(\tau)$ using your answer to problem 3 above.

(c) Create a an image of the kernel $k(s,t)$ for $(s,t) \in [0,1] \times [0,1]$ for the basis above — use a least a few hundred points for each of the arguments s and t. (In MATLAB you can display using `imagesc`.)

5. Let

$$\boldsymbol{A} = \begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 0.98 \end{bmatrix}$$

(a) Find the eigenvalue decomposition of $\boldsymbol{A}$ by hand. Recall that λ is an eigenvalue of $\boldsymbol{A}$ if for some $u[1], u[2]$ (entries of the corresponding eigenvector) we have

$$\begin{aligned}(1.01 - \lambda)u[1] + 0.99u[2] &= 0 \\ .99u[1] + (0.98 - \lambda)u[2] &= 0.\end{aligned}$$

Another way of saying this is that we want the values of λ such that $\boldsymbol{A} - \lambda\mathbf{I}$ (where $\mathbf{I}$ is the 2×2 identity matrix) has a non-trivial null space — there is a nonzero vector $\boldsymbol{u}$ such that $(\boldsymbol{A} - \lambda\mathbf{I})\boldsymbol{u} = 0$. Yet another way of saying this is that we want the values of λ such that $\det(\boldsymbol{A} - \lambda\mathbf{I}) = 0$. Once you have found the two eigenvalues, you can solve the 2×2 systems of equations $\boldsymbol{A}\boldsymbol{u}_1 = \lambda_1\boldsymbol{u}_1$ and $\boldsymbol{A}\boldsymbol{u}_2 = \lambda_2\boldsymbol{u}_2$ for $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$.

Show your work above, but feel free to check you answer using MATLAB/numpy.

(b) If $\boldsymbol{y} = \begin{bmatrix} 1 & 1 \end{bmatrix}^{\mathrm{T}}$, determine the solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$.

(c) Now let $y = \begin{bmatrix} 1.1 & 1 \end{bmatrix}^{\mathrm{T}}$ and solve $Ax = y$. Comment on how the solution changed.

(d) Suppose we observe

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}$$

with $\|\boldsymbol{e}\|_2 = 1$. We form an estimate $\tilde{\boldsymbol{x}} = \boldsymbol{A}^{-1}\boldsymbol{y}$. Which vector $\boldsymbol{e}$ (over all error vectors with $\|\boldsymbol{e}\|_2 = 1$) yields the maximum error $\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_2^2$?

(e) Which (unit) vector $\boldsymbol{e}$ yields the minimum error?

(f) Suppose the components of $\boldsymbol{e}$ are iid Gaussian:

$$e[i] \sim \mathrm{Normal}(0, 1).$$

What is the mean-square error $\mathrm{E}[\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_2^2]$?

(g) Verify your answer to the previous part in MATLAB by taking $\boldsymbol{A}\boldsymbol{x} = \begin{bmatrix} 1 & 1 \end{bmatrix}^{\mathrm{T}}$, and then generating 10,000 different realizations of $\boldsymbol{e}$ using the `randn` command, and then averaging the results. Turn in your code and the results of your computation.

6. (a) Let $\boldsymbol{A}$ be a $N \times N$ symmetric matrix. Show that[1]

$$\mathrm{trace}(\boldsymbol{A}) = \sum_{n=1}^{N} \lambda_n,$$

where the $\{\lambda_n\}$ are the eigenvalues of $\boldsymbol{A}$.

(b) Now let $\boldsymbol{A}$ be an arbitrary $M \times N$ matrix. Recall the definition of the Frobenius norm:

$$\|\boldsymbol{A}\|_F = \left(\sum_{m=1}^{M} \sum_{n=1}^{N} |A[m,n]|^2 \right)^{1/2}.$$

[1] The trace of a (square) matrix is the sum of the elements on the diagonal: $\mathrm{trace}(\boldsymbol{A}) = \sum_{n=1}^{N} A[n,n]$.

Show that

$$\|\boldsymbol{A}\|_F^2 = \operatorname{trace}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}) = \sum_{r=1}^{R} \sigma_r^2,$$

where R is the rank of $\boldsymbol{A}$ and the $\{\sigma_r\}$ are the singular values of $\boldsymbol{A}$.

(c) The *operator norm* (sometimes called the *spectral norm*) of an $M \times N$ matrix is

$$\|\boldsymbol{A}\| = \max_{\boldsymbol{x}\in\mathbb{R}^N,\ \|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2.$$

(This matrix norm is so important, it doesn't even require a designation in its notation — if somebody says "matrix norm" and doesn't elaborate, this is what they mean.) Show that

$$\|\boldsymbol{A}\| = \sigma_1,$$

where σ_1 is the largest singular value of $\boldsymbol{A}$. For which $\boldsymbol{x}$ does

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 = \|\boldsymbol{A}\| \cdot \|\boldsymbol{x}\|_2 \ \ ?$$

(d) Prove that $\|\boldsymbol{A}\| \leq \|\boldsymbol{A}\|_F$. Give an example of an $\boldsymbol{A}$ with $\|\boldsymbol{A}\| = \|\boldsymbol{A}\|_F$.

7. The file `hw4p7.mat` contains two variables: `udata` and `ydata`. We will use this data to estimate a function $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}$. The columns of `udata` contain sample locations, there are $M = 100$. The entries of `y` are the corresponding responses. We want to $\boldsymbol{f}$ such that

$$f(\boldsymbol{u}_m) \approx y_m, \quad m = 1, \ldots, M, \quad \text{where} \quad \boldsymbol{u}_m = \begin{bmatrix} s_m \\ t_m \end{bmatrix}.$$

We will restrict f to be a second-order polynomial on $[0,1] \times [0,1]$:

$$f(s,t) = \alpha_1 s^2 + \alpha_2 t^2 + \alpha_3 st + \alpha_4 s + \alpha_5 t + \alpha_6,$$

which means that $\boldsymbol{f}$ lies is in six dimensional subspace of $L_2([0,1]^2)$.

(a) Explain how to compute the 100×6 matrix $\boldsymbol{A}$ so that $\boldsymbol{y} \approx \boldsymbol{A}\boldsymbol{\alpha}$, where $\boldsymbol{y}$ contains the 100 response values in `ydata`. Write the code to compute $\boldsymbol{A}$ and turn it in.

(b) Solve

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^6}{\text{minimize}} \ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\alpha}\|_2^2.$$

Turn in your code and the numerical value of your solution $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^6$.

(c) Make a contour plot of the corresponding

$$\hat{f}(s,t) = \hat{\alpha}_1 s^2 + \hat{\alpha}_2 t^2 + \hat{\alpha}_3 st + \hat{\alpha}_4 s + \hat{\alpha}_5 t + \hat{\alpha}_6.$$

Include 25 contour lines, just so we have a very clear picture of what this function looks like.