Math Foundations of ML, Fall 2017

Homework #5

Due Friday Ocrober 13, 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. 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 $\boldsymbol{Ax} = \begin{bmatrix}1 & 1\end{bmatrix}^{\mathrm{T}}$, 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.

3. Consider the set of bump basis vectors $\psi_1(t), \ldots, \psi_N(t)$, where

$$\psi_k(t) = g\left(t - k/N\right), \quad g(t) = e^{-200t^2} \tag{1}$$

Given a point t, define the nonlinear "feature map" as

$$\boldsymbol{\Psi}(t) = \begin{bmatrix} \psi_1(t) \\ \psi_2(t) \\ \vdots \\ \psi_N(t) \end{bmatrix}$$

Plot the feature map as a discrete set of coefficients[1] for $t = 1/3$ for $N = 10, 20, 50, 100, 200$. Compare to the radial basis kernel map

$$h_t(s) = k(s,t) = e^{-200|s-t|^2},$$

for $t = 1/3$ and $s \in [0,1]$. Discuss the relationship between kernel regression with a Gaussian radial basis function, and nonlinear regression using a basis of the form (1).

4. In this problem, we will solve a stylized regression problem using the data set `hw5p4data.mat`. This file contains (noisy) samples of a function $f(t)$ for $t \in [0,1]$. The sample locations are in the vector `T`, the sample values are in `y`.

(a) Find the best cubic fit to the data using least-squares. That is, find $w_0, \ldots, w_3$ that minimizes

$$\underset{\boldsymbol{w}}{\text{minimize}} \sum_{m=1}^{M} (y_m - f(t_m))^2 \quad \text{where} \quad f(t) = w_3 t^3 + w_2 t^2 + w_1 t + w_0$$

Let $\hat{\boldsymbol{w}}$ be the solution to the above, and $\hat{f}$ the corresponding cubic polynomial. Compute the *sample error*[2]

$$\left(\sum_{m=1}^{M} (y_m - \hat{f}(t_m))^2 \right)^{1/2} = \|\boldsymbol{y} - \boldsymbol{A}\hat{\boldsymbol{w}}\|_2,$$

where $\boldsymbol{A}$ is the matrix you set up to solve the least-squares problem. Plot your solution $\hat{f}(t)$ for $t \in [0,1]$, and overlay the sample values (t_m, y_m) — the sample values should not have lines connecting them[3].

[1] In MATLAB, use `plot(1:N,Psit(1:N),'o')`.

[2] Also called "training error".

[3] Use `plot(t,y,'o')` in MATLAB, for example.

(b) In this case, the function I happened to use to create the samples is

$$f_{\text{true}}(t) = \frac{\sin(12(t+0.2))}{t+0.2}.$$

Compute the (squared) generalization error

$$\left(\int_0^1 |\hat{f}(t) - f_{\text{true}}(t)|^2 \ \mathrm{d}t\right)^{1/2}.$$

You can either use some numerical integration package to do this (`integral()` in MATLAB), or you can simply sample the functions at 5000 points[4], take the sum of the squared difference, divide by 5000, then take the square root.

(c) Repeat part (a) for polynomials of order $p = 4, 5, 6, 7, 8, 9$ (and so the number of basis functions is $N = 5, 6, 7, 8, 9, 10$). For each experiment, report the largest and smallest singular value of the $\boldsymbol{A}$ matrix. What goes wrong at $p = 8$ and $p = 9$?

(d) For $p = 9$, compute the ridge regression estimate with δ very small, say $\delta = 10^{-6}$. As above, report both the sample error and generalization error and make a plot.

(e) Compute the kernel regression estimate using

$$k(u,t) = e^{-|t-u|^2/2\sigma^2},$$

for $\sigma = \frac{1}{20}, \frac{1}{50}, \frac{1}{100}, \frac{1}{300}, \frac{1}{1000}$. In each of these cases, examine the eigenvalues of the $\boldsymbol{K}$ matrix to come up with a reasonable value of δ to use[5]. As before, produce sample errors, generalization errors, and plots for each case. Make insightful comments about what you are seeing.

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

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

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

(b) Recall the definition of the Frobenius norm of an $M \times N$ matrix:

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

Show that

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

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

[4] 5000 might be overkill here, but these computations are cheap.

[5] One reasonable choice might be the maximum eigenvalue divided by 1000.

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

(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 $\|\boldsymbol{A}\| = \|\boldsymbol{A}\|_F$.