CSE 847 (Spring 2021): Machine Learning— Homework 2

Instructor: Jiayu Zhou

1 Linear Algebra II

1. Compute (by hand) the eigenvalues and the eigenvectors of the following matrix:

$$A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Solution: Let (λ, v) be an eigen pair of the matrix A, then we have

$$\begin{aligned} & Av = \lambda v \\ \Rightarrow & (A - \lambda I_n)v = 0 \\ \Rightarrow & A - \lambda I_n \text{ is singular} \\ \Rightarrow & \det(A - \lambda I_n) = 0 \\ \Rightarrow & \det\left(\begin{bmatrix} 2-\lambda & 1 & 0 \\ 1 & 2-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{bmatrix}\right) = 0 \\ \Rightarrow & \det(A - \lambda I_n) = (2-\lambda)^2(1-\lambda) - (1-\lambda) = 0 \\ \Rightarrow & (\lambda - 1)^2(\lambda - 3) = 0 \end{aligned}$$

Thus the eigenvalues are 1 (with multiplicity 2) and 3. For $\lambda_1 = 1$:

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} v = 0 \Rightarrow v_1 = \begin{bmatrix} \sqrt{2}/2 \\ -\sqrt{2}/2 \\ 0 \end{bmatrix}, v_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

For $\lambda_2 = 3$:

$$\begin{bmatrix} -1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -2 \end{bmatrix} v = 0 \Rightarrow v = \begin{bmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \\ 0 \end{bmatrix}.$$

2. Given the three vectors $v_1 = (2, 0, -1), v_2 = (0, -1, 0)$ and $v_3 = (2, 0, 4)$ in $\mathbb{R}^3$.

   - Show that they form an orthogonal set under the standard Euclidean inner product for $\mathbb{R}^3$ but not an orthonormal set.
   - Turn them into a set of vectors that will form an orthonormal set of vectors under the standard Euclidean inner product for $\mathbb{R}^3$.

   Solution:

   - Show that they form an orthogonal set is equivalent to showing that $v_i^T v_j = 0, \forall i \neq j$.

$$\begin{aligned} v_1^T v_2 &= 2 * 0 + 0 * (-1) + (-1) * 0 = 0 \\ v_1^T v_3 &= 2 * 2 + 0 + (-1) * 4 = 0 \\ v_2^T v_2 &= 0 * 2 + 0 * (-1) + 0 * 4 = 0. \end{aligned}$$

Hence, $\{v_1, v_2, v_3\}$ forms an orthogonal set under the standard Euclidean inner product for $\mathbb{R}^3$. However, note that $\|v_1\|_2 = \sqrt{5} \neq 1$ and $\|v_3\|_2 = 2\sqrt{5} \neq 1$, which means that $\{v_1, v_2, v_3\}$ are not an orthonormal set.

- By normalizing v_1 and v_3, we can obtain the orthonormal set:

$$\left\{ \left(\frac{2}{\sqrt{5}}, 0, -\frac{1}{\sqrt{5}} \right), (0, -1, 0), \left(\frac{1}{\sqrt{5}}, 0, \frac{2}{\sqrt{5}} \right) \right\}$$

3. Suppose that A is an $n \times m$ matrix with linearly independent columns. Show that $A^T A$ is an invertible matrix.

   Solution: Since the rank of $A \in \mathbb{R}^{m \times n}$, the SVD of A is given by

$$A = U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T$$

   where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal, and $\Sigma \in \mathbb{R}^{n \times n}$ is diagonal and nonsingular (no zero diagonal elements). It follows that

$$A^T A = V(\Sigma \; 0) U^T U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T = V \Sigma^2 V^T. \tag{1}$$

   Since V and Σ are nonsingular, $A^T A$ is nonsingular.

4. Suppose that A is an $n \times m$ matrix with linearly independent columns. Let $\bar{x}$ be a least squares solution to the system of equations $Ax = b$ (the solution of $\min_x \|Ax - b\|_2^2$). Show that $\bar{x}$ is the **unique** solution to the associated normal system $A^T A\bar{x} = A^T b$.

   Proof:

   Assume the SVD of A is given by

$$A = (U_1 U_2) \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T,$$

   where $\Sigma \in \mathbb{R}^{n \times n}$ is diagonal and nonsingular, $U_1 \in \mathbb{R}^{m \times n}$ and $U_2 \in \mathbb{R}^{m \times (m-n)}$ have orthonormal columns and $V \in \mathbb{R}^{n \times n}$ is orthogonal. We shown in the class that $\|Ax - b\|_2^2 = \|\Sigma y - b_1\|_2^2 + \|b_2\|_2^2$, where $y = V^T x, b_1 = U_1^T b$ and $b_2 = U_2^T b$. Thus the least squares solution is given by

$$\hat{x} = Vy = V\Sigma^{-1} b_1 = V\Sigma^{-1} U_1^T b$$

   Next, we show that the vector $\hat{x}$ above also satisifies $A^T A\hat{x}$ above satisfies $A^T A\hat{x} = A^T b$. From the derivation for Question 5, we have

$$A^T A\hat{x} = V\Sigma^2 V^T \hat{x} = V\Sigma^2 V^T V\Sigma^{-1} U_1^T b = V\Sigma U_1^T b = A^T b.$$

   Since $A^T A$ is nonsingular (from Question 5), $\hat{x}$ is the unique solution to the normal system.

   Alternative Proof: One alternative proof is via contradiction.

If $\hat{x}$ is not the unique solution the nthere is another solution $\hat{x}' = \hat{x} + \delta x$ be the solution of $A^T A\hat{x}' = A^T b$ with $\delta x \neq 0$.

$$\begin{aligned}
&A^T A\hat{x}' = A^T b \\
\Rightarrow &A^T A(\hat{x} + \delta x) = A^T b \\
\Rightarrow &A^T A\hat{x} + A^T A\delta x = A^T b \\
\Rightarrow &A^T A\delta x = A^T b - A^T A\hat{x} \\
\Rightarrow &\delta x = (A^T A)^{-1}(A^T b - A^T A\hat{x}) \\
&\delta x = (A^T A)^{-1}A^T b - (A^T A)^{-1}A^T A\hat{x} \\
&\delta x = A^{-1}b - \hat{x} = \hat{x} - \hat{x} = 0
\end{aligned}$$

then δx must be zero and $\hat{x}'$ should be equivalent to $\hat{x}$, which means $\hat{x}$ is the unique solution to the system. This thus completes the proof.

2 Linear Regression I

Questions in the textbook Pattern Recognition and Machine Learning:

1. **Page 174, Question 3.2.** Show that the matrix

$$\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$$

takes any vector v and projects it onto the space spanned by the columns of $\mathbf{\Phi}$. Use this result to show that the least-squares solution (3.15):

$$\mathbf{w}_{\mathrm{ML}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t} \tag{3.15}$$

corresponds to an orthogonal projection of the vector $\mathbf{t}$ onto the manifold $\mathcal{S}$ as shown in Figure 3.2.

Proof: We first write

$$\begin{aligned}
\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{v} &= \mathbf{\Phi}\tilde{\mathbf{v}} \\
&= \boldsymbol{\phi}_1\tilde{v}^{(1)} + \boldsymbol{\phi}_2\tilde{v}^{(2)} + \cdots + \boldsymbol{\phi}_M\tilde{v}^{(M)}
\end{aligned}$$

where $\boldsymbol{\phi}_m$ is the m-th column of $\mathbf{\Phi}$ and $\tilde{\mathbf{v}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{v}$. By comparing this with the least squares solution in (3.15), we see that

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w}_{\mathrm{ML}} = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t}$$

corresponds to a projection of $\mathbf{t}$ onto the space spanned by the columns of $\mathbf{\Phi}$. To see that this is indeed an orthogonal projection, we first note that for any column of $\mathbf{\Phi}$, $\boldsymbol{\phi}_j$,

$$\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\boldsymbol{\phi}_j = [\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{\Phi}]_j = \boldsymbol{\phi}_j$$

and therefore

$$(\mathbf{y} - \mathbf{t})^T\boldsymbol{\phi}_j = (\mathbf{\Phi}\mathbf{w}_{\mathrm{ML}} - \mathbf{t})^T\boldsymbol{\phi}_j = \mathbf{t}^T(\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T - \mathbf{I})^T\boldsymbol{\phi}_j = 0$$

and thus $(\mathbf{y} - \mathbf{t})$ is orthogonal to every column of $\mathbf{\Phi}$ and hence is orthogonal to $\mathcal{S}$.

2. **Page 175, Question 3.7.** By using the technique of completing the square, verify the result (3.49) for the posterior distribution of the parameters $\mathbf{w}$ in the linear basis function model in which $\mathbf{m}_N$ and $\mathbf{S}_N$ are defined by (3.50) and (3.51) respectively.

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \tag{3.49}$$
$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{t}) \tag{3.50}$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi} \tag{3.51}$$

Solution: From Bayes' theorem we have

$$p(\mathbf{w}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{w})p(\mathbf{w})$$

where the factors on the r.h.s are given by (3.10) and (3.48), respectively.

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \tag{3.10}$$
$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \tag{3.48}$$

Writing this out in full, we get

$$\begin{aligned}
p(\mathbf{w}|\mathbf{t}) &\propto \left[\prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})\right] \mathcal{N}(\mathbf{w}|_0, \mathbf{S}_0) \\
&\propto \exp\left(-\frac{\beta}{2}(\mathbf{t} - \mathbf{\Phi}\mathbf{w})^T(\mathbf{t} - \mathbf{\Phi}\mathbf{w})\right) \\
&\quad \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0)\right) \\
&= \exp\left(-\frac{1}{2}\left(\mathbf{w}^T(\mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi})\mathbf{w} - \beta\mathbf{t}^T\mathbf{\Phi}\mathbf{w} - \beta\mathbf{t}^T\mathbf{\Phi}\mathbf{w} - \beta\mathbf{w}^T\mathbf{\Phi}^T\mathbf{t} + \beta\mathbf{t}^T\mathbf{t}\right.\right. \\
&\qquad \left.\left. -\mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{w} - \mathbf{w}^T\mathbf{S}_0^{-1}\mathbf{m}_0 + \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\mathbf{w}^T(\mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi})\mathbf{w} - (\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{t})^T\mathbf{w}\right.\right. \\
&\qquad \left.\left. -\mathbf{w}^T(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{t}) + \beta\mathbf{t}^T\mathbf{t} + \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0\right)\right) \\
&= \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^T\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N)\right) \\
&\quad \exp\left(-\frac{1}{2}(\beta\mathbf{t}^T\mathbf{t} + \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0 - \mathbf{m}_N^T\mathbf{S}_N^{-1}\mathbf{m}_N)\right)
\end{aligned}$$

where we have used (3.50) and (3.51) when completing the square in the last step. The first exponential corrsponds to the posterior, unnormalized Gaussian distribution over $\mathbf{w}$, while the second exponential is independent of $\mathbf{w}$ and hence can be absorbed into the normalization factor.

3. **Page 175, Question 3.10.** By making use of the result (2.115) to evaluate the integral in (3.57), verify that the predictive distribution for the Bayesian linear regression model is given by (3.58) in which the input-dependent variance is given by (3.59).

$$p(t|\mathbf{t},\alpha,\beta) = \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)d\mathbf{w} \tag{3.57}$$

$$p(t|\mathbf{x},\mathbf{t},\alpha,\beta) = \mathcal{N}(t|\mathbf{m}_N^T\boldsymbol{\phi}(\mathbf{x}),\sigma_N^2(\mathbf{x})) \tag{3.58}$$

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}). \tag{3.59}$$

Marginal and Conditional Gaussians Given a marginal Gaussian distribution for $\mathbf{x}$ and a conditional Gaussian distri- bution for y given $\mathbf{x}$ in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1}) \tag{2.113}$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b},\mathbf{L}^{-1}) \tag{2.114}$$

the marginal distribution of y and the conditional distribution of x given y are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \tag{2.115}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\},\boldsymbol{\Sigma}) \tag{2.116}$$

where

$$\Sigma = (\boldsymbol{\Lambda}+\mathbf{A}^T\mathbf{L}\mathbf{A})^{-1} \tag{2.117}$$

Solution: Using (3.3), (3.8) and (3.49),

$$y(\mathbf{x},\mathbf{w}) = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}) \tag{3.3}$$

$$p(t|\mathbf{x},\mathbf{w},\beta) = \mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1}) \tag{3.8}$$

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N) \tag{3.49}$$

we can re-write (3.57) as

$$p(t|\mathbf{x},\mathbf{t},\alpha,\beta) = \int p(t|\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)d\mathbf{w} = \int \mathcal{N}(t|\boldsymbol{\phi}(\mathbf{x})^T\mathbf{w},\beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}_N,\mathbf{S}_N)d\mathbf{w}$$

By matching the first factor of the integrand with (2.114) and the second factor with (2.113), we obtain the desired result directly from (2.115).

4. **Page 175, Question 3.11.** We have seen that, as the size of a data set increases, the uncertainty associated with the posterior distribution over model parameters decreases. Make use of the matrix identity (Appendix C)

$$(\mathbf{M}+\mathbf{v}\mathbf{v}^T)^{-1} = \mathbf{M}^{-1} - \frac{(\mathbf{M}^{-1}\mathbf{v})(\mathbf{v}^T\mathbf{M}^{-1})}{1+\mathbf{v}^T\mathbf{M}^{-1}\mathbf{v}} \tag{3.110}$$

to show that the uncertainty $\sigma_N^2(\mathbf{x})$ associated with the linear regression function given by (3.59) satisfies

$$\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x})$$

Solution: From (3.59)

$$\sigma^2_{N+1}(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_{N+1} \boldsymbol{\phi}(\mathbf{x}) \tag{3.59}$$

where $\mathbf{S}_{N+1}$ is given by (r.f. Question 3.8):

$$\mathbf{S}^{-1}_{N+1} = \mathbf{S}^{-1}_N + \beta \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}^T_{N+1}. \tag{P}$$

From (P) and (3.110) we get

$$\begin{aligned}
\mathbf{S}_{N+1} &= (\mathbf{S}^{-1}_N + \beta \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}^T_{N+1})^{-1} \\
&= \mathbf{S}_N - \frac{(\mathbf{S}_N \boldsymbol{\phi}_{N+1} \beta^{1/2})(\beta^{1/2} \boldsymbol{\phi}^T_{N+1} \mathbf{S}_N)}{1 + \beta \boldsymbol{\phi}^T_{N+1} \mathbf{S}_N \boldsymbol{\phi}_{N+1}} \\
&= \mathbf{S}_N - \frac{\beta \mathbf{S}_N \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}^T_{N+1} \mathbf{S}_N}{1 + \beta \boldsymbol{\phi}^T_{N+1} \mathbf{S}_N \boldsymbol{\phi}_{N+1}}
\end{aligned}$$

Using this and (3.59), we can rewrite (3.59) as

$$\begin{aligned}
\sigma^2_{N+1}(\mathbf{x}) &= \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \left(\mathbf{S}_N - \frac{\beta \mathbf{S}_N \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}^T_{N+1} \mathbf{S}_N}{1 + \beta \boldsymbol{\phi}^T_{N+1} \mathbf{S}_N \boldsymbol{\phi}_{N+1}} \right) \boldsymbol{\phi}(\mathbf{x}) \\
&= \sigma^2_N(\mathbf{x}) - \frac{\beta \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}_{N+1} \boldsymbol{\phi}^T_{N+1} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})}{1 + \beta \boldsymbol{\phi}^T_{N+1} \mathbf{S}_N \boldsymbol{\phi}_{N+1}}
\end{aligned}$$

Since $\mathbf{S}_N$ is positive definite, the numerator and denominator of the second term above will be non-negative and positive, respectively, and hence $\sigma^2_{N+1}(\mathbf{x}) \leq \sigma^2_N(\mathbf{x})$.