Homework 3

❗ For each problem, please clearly show your reasoning and write all the steps.
G As data scientists, you should feel free to google it whenever you see something unfamiliar.
☺ Group discussion for the homework is encouraged, but you have to write your answer by yourself. Also, you are always welcome to discuss the problems with me.

Task 0.

Read the relevant chapters in the textbooks listed on Sakai.

- [Al] stands for *Introduction to Machine Learning* by Alpaydin;
- [Bi] stands for *Patern Recognition and Machine Learning* by Bishop;
- [HaTF] stands for *The Elements of Statistical Learning* by Hastie, Tibshirani and Friedman.

Problem 1. Programming: Spectral clustering vs K-means (10pt) cf. [HaTF] Fig. 14.29

In $\mathbb{R}^2$, for each of the three concentric clusters with radius 1, 2.8 and 5, generate 150 points. Add a noise $\sim \mathcal{N}(0, 0.25^2)$ to each of the 450 points to form your dataset.

1. Using K-means with $K = 3$, cluster the data points. Show your results in 2D and in three colors.
2. Construct a similarity graph with k-NN where $k = 10$. Plot the graph Laplacian $\boldsymbol{L}$ ("plt.imshow" your matrix).
3. Construct the feature matrix $\boldsymbol{Z}$ (which is 450×2 using the two eigenvectors of $\boldsymbol{L}$ corresponding to the second and the third smallest eigenvalues. Then, using K-means with $K = 3$, cluster the features. Show your results in 2D for the original data points.

Problem 2. RKHS (10pt) [HaTF] Ex.5.15

Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) for which the kernel satisfies the eigendecomposition

$$K(x, y) = \sum_{i=1}^{\infty} \gamma_i \psi_i(x) \psi_i(y) \ .$$

Recall that

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i=1}^{\infty} \frac{\langle f, \psi_i \rangle \cdot \langle g, \psi_i \rangle}{\gamma_i} \ .$$

Remark: note that $\langle \cdot, \cdot \rangle$ is different from $\langle \cdot, \cdot \rangle_{\mathcal{H}}$.
Prove the following:

1. $\langle K(\cdot, x_i), f\rangle_{\mathcal{H}} = f(x_i)$.
2. $\langle K(\cdot, x_i), K(\cdot, x_j)\rangle_{\mathcal{H}} = K(x_i, x_j)$.
3. If $g(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i)$, then

$$\|g\|_{\mathcal{H}}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} K(x_i, x_j)\alpha_i\alpha_j \ .$$

4. Suppose $\tilde{g}(x) = g(x) + \rho(x)$, with $\rho(x) \in \mathcal{H}$ and orthogonal in $\mathcal{H}$ to each of $K(x, x_i)$ (this means that $\langle \rho, \phi_i\rangle_{\mathcal{H}} = 0$ for all i). Show that

$$\sum_{i=1}^{N} L(y_i, \tilde{g}(x_i)) + \lambda \|\tilde{g}\|_{\mathcal{H}}^2 \geq \sum_{i=1}^{N} L(y_i, g(x_i)) + \lambda \|g\|_{\mathcal{H}}^2 \ .$$

(cf. [Bi] Ex. 6.16)

Problem 3. Gaussian process (10pt+4pt bonus+4pt bonus)

Solve any one of the following three problems. You can choose to solve all three and if you do so, you need to assign two of them as bonus problems and you may get at most 4 bonus points from each bonus problem. If you don't specify your assignment, the first problem will be treated as a "regular problem" and the last two problems will be treated as the bonus problems.

1. (cf. [Bi] Pages 93, 87) Prove the following.
   (a) If $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ and $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\boldsymbol{A}\mathbf{x} + \boldsymbol{b}, \boldsymbol{L}^{-1})$, then $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^{\mathrm{T}})$.
   (b) If $\mathbf{x} = \begin{bmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{bmatrix}$ and $p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x}\middle| \begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}, \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \hline \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{array}\right]\right)$, then

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}) \ ,$$

   where $\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)$ and $\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}$.
2. (cf. [Bi] §6.4.7) Consider a neural network with one hidden layer. Suppose the input dimension is N_0, the hidden dimension is N_1 and the output dimension is 1 (i.e. the output is a scalar). Let's denote the neural network by

$$\begin{aligned} x_j^1 &= \sigma\left(b_j^0 + \sum_{k=1}^{N_0} W_{jk}^0 x_k\right) \qquad \text{for } j = 1, \cdots, N_1 \\ z &= b_1^1 + \sum_{j=1}^{N_1} W_{1j}^1 x_j^1 \end{aligned}$$

   where the parameters are all independent with each $W_{ij}^l \sim$ i.i.d. $\mathcal{N}(0, \sigma_w^2/N_l)$ and $b_i^l \sim$ i.i.d. $\mathcal{N}(0, \sigma_b^2)$, $l = 0, 1$, respectively. Take $N_1 \to \infty$. Prove that $z = z(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, \cdots, x_{N_0})^{\mathrm{T}} \in \mathbb{R}^{N_0}$, converges to a Gaussian process $\mathcal{GP}(\mu, K)$. Specify μ and K.
3. Study [Bi] §6.4.6. Recall that in class we mentioned that $p(\boldsymbol{a}_N|\mathtt{t}_N)$ is approximated by a Gaussian.

(a) In plain language, state what the Gaussian is designed to be.

(b) In plain language, state the strategy you will follow to solve for the parameters of the Gaussian.

(c) Explain in detail how to derive (6.80)-(6.82), and then briefly discuss how to use these results (i.e. briefly discuss what's been done in (6.83-6.88)).

Problem 4. Journal (10pt)

Write a journal about what you have learned during this week. You can think about e.g. the following questions: What have you learned this week? Which topic is the most interesting? Which topic is the most difficult? What application is useful? What techniques have you mastered during the week? What can the instructor do to improve the learning process? What have you learned from your classmates?