CS 189 Introduction to Machine Learning
Fall 2018 DIS3

1 Kernels

For a function $k(\mathbf{x}_i, \mathbf{x}_j)$ to be a valid kernel, it suffices to show either of the following conditions is true:

1. k has an inner product representation: $\exists\ \Phi : \mathbb{R}^d \to \mathcal{H}$, where $\mathcal{H}$ is some (possibly infinite-dimensional) inner product space such that $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d,\ k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$.
2. For every sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in R^d$, the Gram matrix

$$\mathbf{K} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_n) \\ \vdots & k(\mathbf{x}_i, \mathbf{x}_j) & \vdots \\ k(\mathbf{x}_n, \mathbf{x}_1) & \cdots & k(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}$$

is positive semidefinite. For the following parts you can use either condition (1) or (2) in your proofs.

(a) Show that the first condition implies the second one, i.e. if $\forall \mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^d,\ k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ then the Gram matrix K is PSD.

Solution: $\forall \mathbf{a} \in \mathbb{R}^n, \mathbf{a^T K a} = \sum_{i,j} \mathbf{a}_i \mathbf{a}_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i,j} \mathbf{a}_i \mathbf{a}_j \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = \| \sum_i \mathbf{a}_i \Phi(\mathbf{x}_i) \|_2^2 \geq 0$

(b) Given two valid kernels k_a and k_b, show that their linear combination

$$k(\mathbf{x}_i, \mathbf{x}_j) = \alpha k_a(\mathbf{x}_i, \mathbf{x}_j) + \beta k_b(\mathbf{x}_i, \mathbf{x}_j)$$

is a valid kernel, where $\alpha \geq 0$ and $\beta \geq 0$.

Solution: We can show that $\mathbf{K}$ is positive semidefinite: $x^\top \mathbf{K} x = x^\top (\alpha \mathbf{K}_a + \beta \mathbf{K}_b) x = \alpha x^\top \mathbf{K}_a x + \beta x^\top \mathbf{K}_b x \geq 0$

(c) Given a valid kernel k_a, show that

$$k(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i) f(\mathbf{x}_j) k_a(\mathbf{x}_i, \mathbf{x}_j)$$

is a valid kernel.

Solution: We can show k admits a valid inner product representation:

$$k(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i) f(\mathbf{x}_j) \langle \Phi_a(\mathbf{x}_i), \Phi_a(\mathbf{x}_j) \rangle = \langle f(\mathbf{x}_i) \Phi_a(\mathbf{x}_i), f(\mathbf{x}_j) \Phi_a(\mathbf{x}_j) \rangle = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$$

where $\Phi(x) = f(x) \Phi_a(x)$.

(d) Given a positive semidefinite matrix A, show that $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top A \mathbf{x}_j$ is a valid kernel.

Solution: We can show k admits a valid inner product representation:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top A \mathbf{x}_j = \mathbf{x}_i^\top P D^{1/2} D^{1/2} P^\top \mathbf{x}_j = \langle D^{1/2} P^\top \mathbf{x}_i, D^{1/2} P^\top \mathbf{x}_j \rangle = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$$

where $\Phi(x) = D^{1/2} P^\top x$

(e) Show why $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top (\text{rev}(\mathbf{x}_j))$ (where $\text{rev}(x)$ reverses the order of the components in x) is *not* a valid kernel.

Solution: We have that $k((-1,1),(-1,1)) = -2$, but this is invalid since if k is a valid kernel then $\forall \mathbf{x},\ k(\mathbf{x}, \mathbf{x}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}) \rangle \geq 0$.

(f) When solving Kernel ridge regression, one can show that the key intermediate step is solving the following optimization problem:

$$\text{argmin}_{\alpha \in \mathbb{R}^n} \left[\frac{1}{2} \alpha^T (\mathbf{K} + \lambda I) \alpha - \lambda \langle \alpha, y \rangle \right]$$

where $y \in \mathbb{R}^n$, $\lambda \geq 0$, and $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the Gram matrix computed by applying a kernel function k on every sample pair: $k(\mathbf{x}_i, \mathbf{x}_j)$. When λ is close to 0, why is it important that $\mathbf{K}$ is a valid kernel?

Solution: Since $\mathbf{K}$ is a symmetric matrix, it has a spectral of the form: $\mathbf{K} = Q \Lambda Q^\top$. If $\mathbf{K}$ is not a valid kernel function, then there must be a negative eigenvalue. Without the loss of generality, assume that it is $\Lambda_{11} < 0$. Let $\alpha = cQe_1$, where e_1 is the first cannonical basis vector, and $c \in \mathbb{R}$. Then the argmin is at most

$$\begin{aligned}
\frac{1}{2} \alpha^\top ((\mathbf{K} + \lambda I)) \alpha - \lambda \langle \alpha, y \rangle &= \frac{1}{2} (cQe_1)^\top (Q \Lambda Q^\top + \lambda I)(cQe_1) - \lambda \langle cQe_1, y \rangle \\
&= \frac{1}{2} c^2 e_1^\top (Q^\top Q \Lambda Q^\top Q + \lambda Q^\top Q) e_1 - c\lambda \langle Qe_1, y \rangle \\
&= \frac{1}{2} c^2 e_1^\top (\Lambda + \lambda I) e_1 - c\lambda \langle Qe_1, y \rangle \\
&= c^2 \cdot \frac{\Lambda_{11} + \lambda}{2} - c \langle Qe_1, y \rangle.
\end{aligned}$$

When $\lambda < -\Lambda_{11}$, then as $c \to \infty$, the term $c^2 \cdot \frac{\Lambda_{11} + \lambda}{2}$ dominates $c\langle Qe_1, y \rangle$, and tends to $-\infty$. Therefore, the arg min does not exist if $\mathbf{K}$ has a strictly negative eigenvalue (and $\lambda < -\Lambda_{11}$).

2 Multivariate Gaussians: A review

(a) Consider a two dimensional random variable $Z \in \mathbb{R}^2$. In order for the random variable to be jointly Gaussian, a necessary and sufficient condition is that

- Z_1 and Z_2 are each marginally Gaussian, and

- $Z_1|Z_2 = z$ is Gaussian, and $Z_2|Z_1 = z$ is Gaussian.

A second characterization of a jointly Gaussian RV Z is that it can be written as $Z = AX$, where X is a collection of i.i.d. standard normal RVs and $A \in \mathbb{R}^{2\times 2}$ is a matrix.

Note that the probability density function of a Gaussian RV is:

$$f(z) = exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)/\sqrt{(2\pi)^k|\Sigma|}$$

.

Let X_1 and X_2 be i.i.d. standard normal RVs. Let U denote a random variable uniformly distributed on $\{-1, 1\}$, independent of everything else. Verify if the conditions of the first characterization hold for the following random variables, and calculate the covariance matrix Σ_Z.

- $Z_1 = X_1$ and $Z_2 = X_2$.
- $Z_1 = X_1$ and $Z_2 = X_1 + X_2$. (Use the second characterization to argue joint Gaussianity.)
- $Z_1 = X_1$ and $Z_2 = -X_1$.
- $Z_1 = X_1$ and $Z_2 = UX_1$.

Solution: Before diving into the solution, recall that the covariance matrix of a vector random variable X with mean (vector) μ is given by $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^\top]$. In other words, entry i, j of the covariance matrix denotes the covariance between the random variables X_i and X_j, i.e., $\Sigma_{ij} = \mathsf{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$.

Additionally, two random variables U and V are said to be uncorrelated if $\mathsf{cov}(U, V) = 0$

- Z_1 and Z_2 are i.i.d. standard Gaussian, and so $(Z_1|Z_2 = z) \sim N(0, 1)$. Also, $Z_2|Z_1 = z \sim N(0, 1)$. Hence, the RVs are jointly Gaussian. We also have $\Sigma_Z = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
- $Z_1 \sim N(0, 1)$, and $Z_2 \sim N(0, 2)$, but these RVs are not independent. Also, we have $(Z_2|Z_1 = z) \sim N(z, 1)$. In order to calculate the distribution of $(Z_1|Z_2 = z)$, see part (e).

  Using the second characterization of joint Gaussianity, it is clear that Z is jointly Gaussian, with $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$. The covariance matrix is given by $\Sigma_Z = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$.
- We have $Z_1 \sim N(0, 1)$ and $Z_2 \sim N(0, 1)$ marginally. However, we have $(Z_1|Z_2 = z) \sim N(-z, 0)$, which is a degenerate Gaussian. The other conditional distribution is identical. Hence, the RVs are jointly Gaussian. The covariance matrix is given by $\Sigma_Z = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$.
- As before, we have $Z_1 \sim N(0, 1)$ and $Z_2 \sim N(0, 1)$ marginally. In order to see this, write

$$p(Z_2 = z_2) = p(Z_2 = z_2|U = 1)p(U = 1) + p(Z_2 = z_2|U = -1)p(U = -1)$$

$$
\begin{aligned}
&= \frac{1}{2}p(X_1 = z_2|U = 1) + \frac{1}{2}p(X_1 = -z_2|U = -1) \\
&= \frac{1}{2}p(X_1 = z_2) + \frac{1}{2}p(X_1 = -z_2) \\
&= \frac{1}{2}p(X_1 = z_2) + \frac{1}{2}p(X_1 = z_2) \\
&= p(X_1 = z_2)
\end{aligned}
$$

The random variable $(Z_2|Z_1 = z)$ is uniformly distributed on $\{-z, z\}$, and is therefore not Gaussian. The RVs are therefore not jointly Gaussian. The covariance matrix is given by $\Sigma_Z = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

(b) Use the above example to show that two Gaussian random variables can be uncorrelated, but not independent. On the other hand, show that two uncorrelated, jointly Gaussian RVs are independent.

Solution: By definition, two random variables X and Y are independent iff $P(X = x, Y = y) = P(X = x)P(Y = y), \forall x, y$.

The last example in the previous part shows uncorrelated Gaussians that are not independent. In order to show that jointly Gaussian RVs (with individual variances σ_1^2 and σ_2^2) that are uncorrelated are also independent, assume without loss of generality that the RVs have zero mean, and notice that one can write the joint pdf as

$$
\begin{aligned}
f_Z(z_1, z_2) &= \frac{1}{(2\pi)\det\left(\Sigma_Z^{1/2}\right)} \exp\left(-\frac{1}{2}\begin{bmatrix} z_1 & z_2 \end{bmatrix}(\Sigma_Z)^{-1}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}\right) \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}}\frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2\sigma_1^2}z_1^2\right)\exp\left(-\frac{1}{2\sigma_1^2}z_2^2\right) \\
&= f_{Z_1}(z_1)f_{Z_2}(z_2).
\end{aligned}
$$

The decomposition follows since Σ_Z is a diagonal matrix when the RVs are uncorrelated. Since we have expressed the joint PDF as a product of the individual PDFs, the RVs are independent.

(c) With the setup above, let $Z = VX$, where $V \in \mathbb{R}^{2\times 2}$, and $Z, X \in \mathbb{R}^2$. What is the covariance matrix Σ_Z? Is this also true for a RV other than Gaussian?

Solution: The covariance matrix of a random vector Z (by definition) is given by $\mathbb{E}(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^\top$. Since the mean $\mathbb{E}[Z]$ is 0, we may write $\Sigma_Z = \mathbb{E}[VXX^\top V^\top] = V\mathbb{E}[XX^\top]V^\top = VV^\top$. This follows by linearity of expectation applied to vector random variables (write it out to convince yourself!)

Yes, this relation is also true for other distributions, since we didn't use the Gaussian assumption in this proof.

(d) Use the above setup to show that $X_1 + X_2$ and $X_1 - X_2$ are independent. Give another example pair of linear combinations that are independent.

Solution: By our previous arguments, it is sufficient to show that these are uncorrelated. Calculating the covariance matrix, we have $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, which is diagonal. Any linear combination $Z = VX$ with $VV^\top = D$ for a diagonal matrix D results in uncorrelated random variables.

(e) Given a jointly Gaussian RV $Z \in \mathbb{R}^2$ with covariance matrix $\Sigma_Z = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}$, how would you derive the distribution of $Z_1 | Z_2 = z$?

Hint: The following identity may be useful

$$\begin{bmatrix} a & b \\ b & c \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{b}{c} & 1 \end{bmatrix} \begin{bmatrix} \left(a - \frac{b^2}{c}\right)^{-1} & 0 \\ 0 & \frac{1}{c} \end{bmatrix} \begin{bmatrix} 1 & -\frac{b}{c} \\ 0 & 1 \end{bmatrix}.$$

Solution: One can do this from first principles, by manipulating the densities themselves. However, we will show a linear algebraic method to derive the density. Using the hint, we begin by writing

$$\Sigma^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{\Sigma_{12}}{\Sigma_{22}} & 1 \end{bmatrix} \begin{bmatrix} \left(\Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}}\right)^{-1} & 0 \\ 0 & \frac{1}{\Sigma_{22}} \end{bmatrix} \begin{bmatrix} 1 & -\frac{\Sigma_{12}}{\Sigma_{22}} \\ 0 & 1 \end{bmatrix}.$$

We can now plug this into the density function. Recall that

$$\begin{aligned} f_{Z_1,Z_2}(z_1, z_2) &\propto \exp\left(-\frac{1}{2} \begin{bmatrix} z_1 & z_2 \end{bmatrix} \Sigma^{-1} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \right) \\ &\propto \exp\left(-\frac{1}{2} \begin{bmatrix} z_1 & z_2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -\frac{\Sigma_{12}}{\Sigma_{22}} & 1 \end{bmatrix} \begin{bmatrix} \left(\Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}}\right)^{-1} & 0 \\ 0 & \frac{1}{\Sigma_{22}} \end{bmatrix} \begin{bmatrix} 1 & -\frac{\Sigma_{12}}{\Sigma_{22}} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \right) \\ &\propto \exp\left(-\frac{1}{2} \begin{bmatrix} z_1 - \frac{\Sigma_{12}}{\Sigma_{22}} z_2 & z_2 \end{bmatrix} \begin{bmatrix} \left(\Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}}\right)^{-1} & 0 \\ 0 & \frac{1}{\Sigma_{22}} \end{bmatrix} \begin{bmatrix} z_1 - \frac{\Sigma_{12}}{\Sigma_{22}} z_2 \\ z_2 \end{bmatrix} \right). \end{aligned}$$

Now see that since the square matrix is diagonal, our density decomposes to yield

$$f_{Z_1,Z_2}(z_1, z_2) \propto \exp\left(-\frac{1}{2} (z_1 - \frac{\Sigma_{12}}{\Sigma_{22}} z_2)^2 \left(\Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}}\right)^{-1} \right) \exp\left(-\frac{1}{2\Sigma_{22}} z_2^2 \right).$$

Conditional on $Z_2 = z_2$, we see that $Z_1 | Z_2 = z_2 \sim N(\frac{\Sigma_{12}}{\Sigma_{22}} z_2, \Sigma_{11} - \frac{\Sigma_{12}^2}{\Sigma_{22}})$.