AIMS African Masters in Machine Intelligence 2022

Kernel Methods - Quizz #2 - Solutions

July 1st, 2022

Exercise 1: Practical questions

1. How would you use Ridge regression to perform binary classification?

   Use a function such as the sigmoid to change the output to a probability.

2. (a) What is the kernel trick?

   In models which depend only on inner products between data points, the kernel trick consists in replacing these inner products $x^\top x'$ by a kernel $k(x, x')$. This is in fact implicitly mapping the data points x to another space $\Phi(x)$.

   (b) When can it be useful?

   This can be useful to construct potentially complex non-linear classifiers from linear models which are simple and easy to optimize.

3. What is the disadvantage of using kernel methods when the number of data points is large?

   Model training depends on computing the kernel Gram matrix between the training points, which is of size N^2. For larger N this is very expensive to compute, and it may not even fit in memory.

   Methods such as low-rank approximations exist to alleviate this problem

4. We denote k_σ the RBF kernel with bandwidth σ : $k_\sigma(x, y) = \exp\left(\frac{\|x-y\|^2}{2\sigma^2}\right)$

   (a) What does the Gram matrix K_σ (or the function k_σ) converge to for very small σ? For very large σ?

   $K_\sigma \to I_n$ *if* $\sigma \to 0$ *and* $K_\sigma \to \mathbf{1}$ *if* $\sigma \to +\infty$

   (b) If we were to use k_σ in our model, what would be the effect of these two cases (σ very small or very large) on model performance ?

   σ too small would lead to a model that is too discriminative: $\sigma = 0$ gives a prediction function that is 0 everywhere except on the training points $\to$ *overfitting.*

   If σ is too small the model will not discriminate enough: in the limit where $\sigma \to \infty$ the prediction function converges to a constant $\to$ *underfitting.*

   (c) How would you choose an appropriate value of σ?

   One can start with the median heuristic to get a reasonable value, and do hyper-parameter tuning from there: grid-search, cross-validation etc...

Exercise 2: Kernel basics

Reminders: *Trigonometry*

- $\cos(-x) = \cos(x), \quad \sin(-x) = -\sin(x)$
- $\cos(a - b) = \cos(a)\cos(b) + \sin(a)\sin(b)$

Questions:

1. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Give 3 equivalent necessary and sufficient conditions for k to be a kernel:
   (a) As condition(s) on function k
      - k must be symmetric
      - $\forall N \in \mathbb{N}, a \in \mathbb{R}^N, x \in \mathcal{X}^N, \quad \sum_{i,j=1}^{N} a_i a_j k(x_i, x_j) \geq 0$

   (b) introducing a feature map Φ
   There exists a Hilbert space $\mathcal{F}$ and a mapping $\Phi : \mathcal{X} \to \mathcal{F}$ such that $k(x, x') =< \Phi(x), \Phi(x') >_{\mathcal{F}}$

   (c) using the Gram-matrix K
   $\forall N \in \mathbb{N}, x \in \mathcal{X}^N$, the Gram matrix K is symmetric positive semi-definite

2. State whether the following functions are p.d. kernels or not. Prove your answer.
   (a) $\mathcal{X} =$ anything, $k(x, x') = C$, where $C \in \mathbb{R}_+$ is a constant.
   Choose $\Phi(x) = \sqrt{C}$:
   $$\Phi(x)^\top \Phi(x') = C = k(x, x')$$
   Hence by definition 1(b), k is a p.d. kernel.

   (b) $\mathcal{X} = \mathbb{R}$, $k(x, x') = \cos(x - x')$
   Choose $\Phi(x) = (\cos(x), \sin(x))^\top$:
   $$\Phi(x)^\top \Phi(x') = \cos(x)\cos(x') + \sin(x)\sin(x') = k(x, x')$$
   Hence by definition 1(b), k is a p.d. kernel.

   (c) $\mathcal{X} = \mathbb{R}$, $k(x, x') = \sin(x - x')$
   $k(x', x) = -k(x, x')$: k is not symmetric and therefore is not a p.d. kernel

3. Let K be a p.d. kernel on $\mathcal{X}$, and $\Phi : \mathcal{X} \to \mathcal{F}$ a mapping to a Hilbert space $\mathcal{F}$ (i.e., a "feature space") such that
   $$\forall x, x' \in \mathcal{X}, \quad K(x, x') = \Phi(x)^\top \Phi(x')$$
   . Show that the distance between points in the feature space $\|\Phi(x) - \Phi(x')\|$ only depends on K (Φ must not appear).

   $$\|\Phi(x) - \Phi(x')\|^2 = K(x, x) + K(x', x') - 2K(x, x')$$

Exercise 3: Positional embeddings

Note: *Results from exercise 2 can be reused here*

Questions:

1. Consider the binary classification problem illustrated in figure 1, where the data points x are real numbers, and the labels y are 1 if the integral part (or *floor*) of x is even, and 0 otherwise ($y = \lfloor x \rfloor \mod 2$).

   Find a mapping $\Phi : \mathbb{R} \to \mathcal{F}$ such that the two classes are linearly separable in the feature space $\mathcal{F}$. You may illustrate your answer if you like.

   There were many possible mappings here, such as $\Phi(x) = \lfloor x \rfloor \mod 2$ or $\Phi(x) = \sin(\pi x)$ for instance.

   One such mapping is illustrated in figure 2

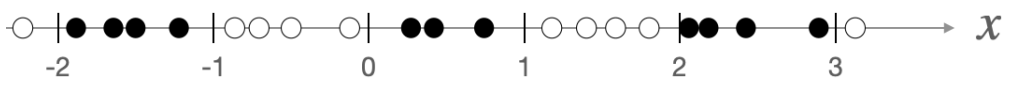

Figure 1: Data distribution for Exercise 3, question 1

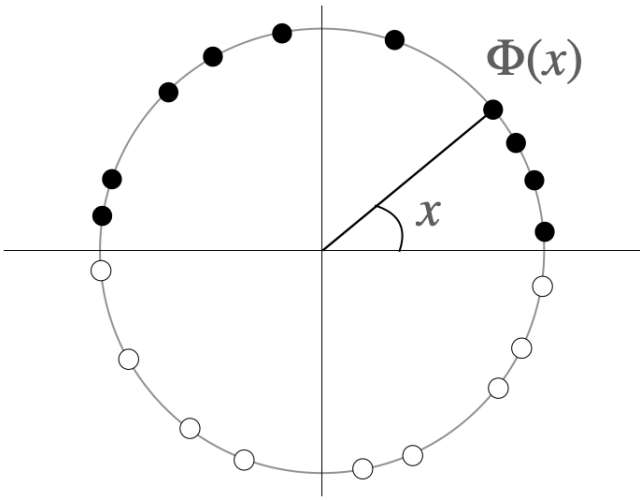

Figure 2: Example $\Phi(x) = (\sin(\pi x), \cos(\pi x))^\top$ that linearly separates the data in Exercise 3, question 1

2. Let $d \in \mathbb{N}$, and $(\omega_i)_{i=1...d} \in \mathbb{R}$, and define

$$K(x, x') = \cos(\omega_1(x - x')) + ... + \cos(\omega_d(x - x')) = \sum_{i=1}^{d} \cos(\omega_i(x - x'))$$

(a) Prove that K is a p.d. kernel, and find a mapping Φ and a feature space $\mathcal{F}$ such that $K(x, x') = \Phi(x)^\top \Phi(x')$

Let

$$\Phi(x) = (\cos(\omega_1 x), \sin(\omega_1 x)..., \cos(\omega_d x), \sin(\omega_d x))^\top$$

Re-using the formula $\cos(a - b) = \cos(a)\cos(b) + \sin(a)\sin(b)$, one can show that:

$$\Phi(x)^\top \Phi(x') = K(x, x')$$

(b) Prove that the norm of Φ is bounded: $\forall x \in \mathbb{R}$, $\|\Phi(x)\| \leq 2d$

In fact $\|\Phi(x)\| = \sqrt{d}$:

$$\|\Phi(x)\|^2 = K(x, x) = \sum_{i=1}^{d} \cos(0) = d$$

(c) Prove that the distance between points in the feature space $\|\Phi(x) - \Phi(x')\|$ only depends on the relative distances $|x - x'|$.

Notice that $K(x, x')$ only depends on $|x - x'|$, and write:

$$\|\Phi(x) - \Phi(x')\|^2 = K(x, x) + K(x', x') - 2K(x, x') = 2d - 2K(x, x')$$

$$\|\Phi(x) - \Phi(x')\|^2 = \sum_{i=1}^{d} (1 - \cos(\omega_i(x - x')))$$

(d) Transformer models, as introduced by [Vaswani et al., 2017] use Φ as a positional encoding for sequences, with $\omega_i = 1/10000^{i/d}$.

Show that, with this choice of frequencies, the indices of a sequence have a unique embedding, i.e. for $n, m \in \mathbb{N}$, $\Phi(m) \neq \Phi(n)$ if $m \neq n$.

$$\|\Phi(m) - \Phi(n)\|^2 = \sum_{i=1}^{d} (1 - \cos(\omega_i(m-n)))$$

Each term in the sum $1 - \cos(\omega_i(m-n))$ is > 0, unless $\omega_i(m-n) = \frac{\pi}{2} \mod \pi$, which cannot happen since $\omega_i(m-n)$ is a rational number.

(e) Give some reasons why Φ was a good choice as a positional embedding.

Φ is a good choice as it is

- of constant norm, hence can scale to sequences of different lengths.
- similarity between two positions $\Phi(m)^\top \Phi(n) = K(m,n)$ only depends on the offset $m - n$
- It separates well the first $\sim$ 10000 integers with only a few dimensions ($d = 512$ in the original model)

For more reasons, see the original article or this blog post:

https://kazemnejad.com/blog/transformer_architecture_positional_encoding/

Bonus exercise: RKHS

What is the RKHS of the linear kernel on $\mathbb{R}^d$? Give the space $\mathcal{H}$ and its inner product. Prove your answer.

This was in the slides (211)

Best wishes, Bonne chance, Yalna yombu !