Maths for Intelligent Systems
Exercise 3

Marc Toussaint, Andrea Baisero
Machine Learning & Robotics lab, U Stuttgart
Universitätsstraße 38, 70569 Stuttgart, Germany

November 20, 2018

1 Eigenvectors

(a) A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is called positive semidefinite (PSD) if $x^\top Ax \geq 0, \forall x \in \mathbb{R}^n$. (PSD is usually only used with symmetric matrices.) Show that *all* eigenvalues of a PSD matrix are non-negative.

(b) Show that if v is an eigenvector of A with eigenvalue λ, then v is also an eigenvector of A^k for any positive integer k. What is the corresponding eigenvalue?

(c) Let v be an eigenvector of A with eigenvalue λ and w an eigenvector of $A^\top$ with a different eigenvalue $\mu \neq \lambda$. Show that v and w are orthogonal with respect to the dot product.

(d) Suppose $A \in \mathbb{R}^{n\times n}$ has eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{R}$. What are the eigenvalues of $A + \alpha I$ for $\alpha \in \mathbb{R}$ and I an identity matrix?

(e) Assume $A \in \mathbb{R}^{n\times n}$ is diagonalizable, i.e., it has n linearly independent eigenvectors, each with a different eigenvalue. Initialize $x \in \mathbb{R}^n$ as a random normalized vector and iterate the two steps

$$x \leftarrow Ax \ , \quad x \leftarrow \frac{1}{\|x\|}\, x$$

Prove that (under certain conditions) these iterations converge to the eigenvector x with a largest (in *absolute* terms $|\lambda_i|$) eigenvalue of A. How fast does this converge? In what sense does it converge if the largest eigenvalue is negative? What if eigenvalues are not different? Other convergence conditions?

(f) Let A be a positive definite matrix with $\lambda_{\max}$ its largest eigenvalue (in absolute terms $|\lambda_i|$). What do we get when we apply power iteration method to the matrix $B = A - \lambda_{max}\mathbf{I}$? How can we get the smallest eigenvalue of A?

(g) Consider the following variant of the previous power iteration:

$$z \leftarrow Ax \ , \quad \lambda \leftarrow x^\top z \ , \quad y \leftarrow (\lambda I - A)y \ , \quad x \leftarrow \frac{1}{\|z\|}\, z \ , \quad y \leftarrow \frac{1}{\|y\|}\, y \ .$$

If A is a positive definite matrix, show that the algorithm can give an estimate of the smallest eigenvalue of A.

2 RKHS

In machine learning we often work in spaces of functions called Reproducing Kernel Hilbert Spaces. These spaces are constructed from a certain type of function called the kernel. The kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ takes two d-dimensional inputs $k(x, x')$, and from the kernel we construct a basis for the space of function, namely $B = \{k(x, \cdot)\}_{x\in\mathbb{R}^d}$. Note that this is a set of infinite element: each $x \in \mathbb{R}^d$ adds a basis function $k(x, \cdot)$ to the basis B. The scalar product between two basis functions $k_x = k(x, \cdot)$ and $k_{x'} = k(x', \cdot)$ is defined to be the kernel evaluation itself: $\langle k_x, k_{x'}\rangle = k(x, x')$. The kernel function is therefore required to be a positive definite function so that it defines a viable scalar product.

(a) Show that for any function $f \in \operatorname{span} B$ it holds

$$\langle f, k_x \rangle = f(x)$$

(b) Assume we only have a finite set of points $\{D = \{x_i\}_{i=1}^n\}$, which defines a finite basis $\{k_{x_i}\}_{i=1}^n \subset B$. This finite function basis spans a subspace $\mathcal{F}_D = \operatorname{span}\{k_{x_i} : x_i \in D\}$ of the space of all functions.

For a general function f, we decompose it $f = f_s + f_\perp$ with $f_s \in \mathcal{F}_D$ and $\forall g \in \mathcal{F}_D : \langle f_\perp, g \rangle = 0$, i.e., $f_\perp$ is orthogonal to $\mathcal{F}_D$. Show that for every $x_i \in D$:

$$f(x_i) = f_s(x_i)$$

(Note: This shows that the function values of any function f at the data points D only depend on the part f_s which is inside the spann of $\{k_{x_i} : x_i \in D\}$. This implies the so-called representer theorem, which is fundamental in kernel machines: A loss can only depend on function values $f(x_i)$ at data points, and therefore on f_s. The part $f_\perp$ can only increase the complexity (norm) of a function. Therefore, the simplest function to optimize any loss will have $f_\perp = 0$ and be within $\operatorname{span}\{k_{x_i} : x_i \in D\}$.)

(c) Within $\operatorname{span}\{k_{x_i} : x_i \in D\}$, what is the coordinate representation of the scalar product?