This is the danger environment.

7.1 Announcements

HW #2: due Monday February 23.

7.2 Outline

- Mercer’s characterization
- Kernel PCA (dimensionality reduction)

7.3 Mercer’s characterization

Given a symmetric and positive semidefinite matrix $K \in \mathbb{R}^{d \times d}$, we know from standard linear algebra that there exist real scalars $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ and vectors $\{\psi_i, i = 1, \cdots, d\}$ such that

$$K = \sum_{i=1}^{d} \lambda_i \psi_i \psi_i^T$$

In this decomposition, the vectors $\{\psi_i\}$ are eigenvectors, obtained by solving the matrix-vector equation

$$K\psi = \lambda\psi.$$

Moreover, the $\{\psi_i\}$ can be chosen to be an orthonormal system of vectors.

We now discuss a generalization of this type of decomposition to the more general setting of linear operators in a Hilbert space. (The matrix is a special case of a linear operator on $\mathbb{R}^d$.)

Given a Hilbert space $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R}\}$ of functions, a linear operator $T : \mathcal{H} \to \mathcal{H}$ is a mapping such that

1. $\forall f \in \mathcal{H}, T(f) \in \mathcal{H}$
2. $\forall f, g \in \mathcal{H}, T(f + g) = T(f) + T(g)$
3. $\forall \alpha \in \mathbb{R}, T(\alpha f) = \alpha T(f)$
7.3.1 Mercer’s theorem (one variant):

**Theorem 7.1.** Say $\mathcal{X} \subseteq \mathbb{R}^d$ is compact, and $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is continuous, and satisfies

$$\int_y \int_x K^2(x,y)dxdy < +\infty,$$

$$\int_y \int_x f(x)K(x,y)f(y)dxdy \geq 0 \ \forall f \in L^2(\mathcal{X}) \ \text{(i.e. a positive semidefinite kernel)}$$

where $L^2(\mathcal{X}) = \{f : \int f^2(x)dx < +\infty\}$

Then there exist $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots$ (all non-negative) and functions $\{\psi_i(\cdot) \in L^2(\mathcal{X}), i = 1, 2, 3, \cdots\}$ such that

$$K(x,y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(y) \ \forall x, y \in \mathcal{X}$$

Moreover, the $\{\psi_i\}$ are an orthonormal system in $L^2(\mathcal{X})$, meaning that

$$\langle \psi_i, \psi_j \rangle_{L^2(\mathcal{X})} = \int \psi_i(x)\psi_j(x)dx = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

**Remarks:** Note that this can be seen as a generalization of the decomposition

$$K(x,y) = x^TKy = \sum_{i=1}^{d} \lambda_i (\psi_i^T x)(\psi_i^T y)$$

in the finite-dimensional setting. The orthogonality condition is a generalization of the fact that PSD matrices have an orthogonal set of eigenvectors.

Mercer’s theorem is a special case of spectral decomposition theory for self-adjoint, positive operators in Hilbert spaces.

7.3.2 Use of Mercer’s Theorem

Eigenfunctions can be obtained by solving the integral equation:

$$T_K(f)(x) := \int \mathbb{K}(x,y)f(y)dy = \lambda f(x)$$

Here

$$T_K(f)(\cdot) := \int \mathbb{K}(\cdot,y)f(y)dy$$

is a linear operator on $L^2(\mathcal{X}) \to L^2(\mathcal{X})$. (Homework #2 has some instances of this procedure.)
We can then use the eigenfunctions thus obtained to generate a “feature map” given by

\[ \Phi : \mathcal{X} \to \ell^2(\mathbb{N}) \]

Here, the feature map \( \Phi \) maps data \( x \in \mathcal{X} \) to a sequence \( (a_1, a_2, \cdots) \in \ell^2(\mathbb{N}) \), where

\[ \ell^2(\mathbb{N}) = \{ (a_1, a_2, \cdots) | \sum_{i=1}^{\infty} a_i^2 < +\infty \} \]

For example, consider the feature map defined as follows:

\[ \Phi(x) = (\sqrt{\lambda_1} \psi_1(x), \sqrt{\lambda_2} \psi_2(x), \cdots, \sqrt{\lambda_i} \psi_i(x), \cdots). \]

That is, we map each \( x \in \mathcal{X} \) into a sequence \( \Phi(x) \) in \( \ell^2(\mathbb{N}) \).

Using Mercer’s decomposition, if we take the inner product (in \( \ell^2(\mathbb{N}) \)) between the two sequences \( \Phi(x) \) and \( \Phi(y) \), then we recover the kernel function

\[ \langle \Phi(x), \Phi(y) \rangle_{\ell^2(\mathbb{N})} = \sum_{i=1}^{\infty} \sqrt{\lambda_i} \psi_i(x) \sqrt{\lambda_i} \psi_i(y) = K(x,y). \]

### 7.4 Kernel PCA

#### 7.4.1 Quick recap on classical PCA

Given data \( X^{(1)}, \cdots, X^{(n)} \subseteq \mathbb{R}^d \), we first compute the sample covariance or correlation matrix, given by

\[ \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X^{(i)}[X^{(i)}]^T \]

Then, we compute the eigenvectors corresponding to the top \( k \ll d \) eigenvalues (in value). Using these eigenvectors, we project data \( X \in \mathbb{R}^d \), a large space, into \( \mathbb{R}^k \), a much smaller space. Thus, the primary motivation for PCA is achieving a large reduction in the dimensionality of the data.

To gain some intuition for PCA, consider an idealized ”noisy subspace” generative model, given by

\[ x = Vz + w \]

where \( V \in \mathbb{R}^{d \times k} \) is fixed, \( z \in \mathbb{R}^d \) is random, and also \( w \in \mathbb{R}^d \) is random. Furthermore, we assume that

\[ \mathbb{E}(z) = 0, \ \text{Cov}(z) = \alpha^2 I_{k \times k} \]
\[ \mathbb{E}(w) = 0, \ \text{Cov}(w) = \sigma^2 I_{d \times d} \]

Finally, we assume that \( z \) and \( w \) are independent. This gives us

\[ \text{Cov}(x) = \Sigma = \alpha^2 VV^T + \sigma^2 I_{d \times d} \]
Now, we may think of \( V \) as having \( k \) orthogonal columns, i.e.,
\[
V = (V_1, \cdots, V_k)
\]
We also have that 
\[
\Sigma V_j = (\alpha^2 + \sigma^2)V_j
\]
i.e., the eigenvectors corresponding to the top \( k \) eigenvalues are \( \{V_1, \cdots, V_k\} \). Moreover, for fixed \( d \), we have that
\[
\|\hat{\Sigma} - \Sigma\|_2 = \max_{\|u\|_2=1} \|(\hat{\Sigma} - \Sigma)u\|_2 \to 0 \text{ as } n \to +\infty
\]
where \( \| \cdot \|_2 \) denotes the spectral radius (max. absolute value over all eigenvalues).

### 7.4.2 Kernel PCA (Scholkopf et. al., 1997)

We once again consider an idealized model, this time in feature space \( \mathcal{F} \), which is given by
\[
\Phi(x) = \sum_{j=1}^{k} z_j \Phi_j + w \tag{7.1}
\]
where \( \Phi_j \in \mathcal{F} \) for all \( j = 1, \cdots, k \) and is fixed, while \( z \in \mathbb{R}^k \) and \( w \in \mathcal{F} \) are both random.

**Example:** Suppose that we worked with the feature map defined by a polynomial kernel \( K(x, y) = (1 + \langle x, y \rangle)^m \) for \( x \in \mathbb{R}^d \). In the special case \( m = 2 \) and \( d = 2 \), one feature map for this kernel is given by
\[
\Phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2)
\]
so that
\[
\langle \Phi(x), \Phi(y) \rangle = 1 + 2x_1y_1 + 2x_2y_2 + 2x_1x_2y_1y_2 + x_1^2 + y_1^2 + x_2^2y_2^2 = (1 + x_1y_1 + x_2y_2)^2
\]
One particular example of the model (7.1) would be
\[
\begin{bmatrix}
1 \\
\sqrt{2}x_1 \\
\sqrt{2}x_2 \\
\sqrt{2}x_1x_2 \\
x_1^2 \\
x_2^2
\end{bmatrix} = z_1 \Phi_1 + w.
\]
This would model the data as lying near to some quadratic surface, determined by the choice of \( \Phi_1 \in \mathbb{R}^6 \).

For simplicity, let us assume that the generating vectors are orthonormal
\[
\langle \Phi_i, \Phi_j \rangle_{\mathcal{F}} = 0 \text{ if } i \neq j
\]
Now let us define the covariance operator associated with the random element $\Phi(x)$. For each $j$, we use $\Phi_j \otimes \Phi_j$ to denote a linear operator on $\mathcal{F}$ defined as follows: given some $f \in \mathcal{F}$, it outputs a new $(\Phi_j \otimes \Phi_j)(f) \in \mathcal{F}$, given by

$$(\Phi_j \otimes \Phi_j)(f) = \langle \Phi_j, f \rangle_{\mathcal{F}} \Phi_j.$$ 

With this definition, the covariance operator is given by

$$\text{Cov}[\Phi(x)] = \sum_{j=1}^{k} \text{Var}(z_j) (\Phi_j \otimes \Phi_j) + \mathbb{E}[w \otimes w]$$

Since it is a linear combination of linear operators, it is also a linear operator on $\mathcal{F}$.

In particular, for any $f \in \mathcal{F}$, this covariance operator outputs a new element of $\mathcal{F}$, given by

$$\text{Cov}[\Phi(x)](f) = \sum_{j=1}^{k} \text{Var}(z_j) \langle \Phi_j, f \rangle_{\mathcal{F}} \Phi_j + \mathbb{E}[w \otimes w](f)$$

At this point, the intuition underlying KPCA is the same as the intuition underlying PCA. That is, if we knew the functions $\Phi_j$, then given a new sample, we could:

- map it to the feature space via $x \mapsto \Phi(x)$
- compute its co-ordinates in the linear span of $\{\Phi_j\}$ by computing the projections $\langle \Phi(x), \Phi(x) \rangle_{\mathcal{F}}$ for $j = 1, \ldots, k$.

In practice, we don’t know the $\{\Phi_j\}$, but as with ordinary PCA, we can try to estimate them from data. Given samples $x^{(i)}, i = 1, 2, \ldots, n$, we can form the empirical covariance operator

$$\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^{n} \Phi(x^{(i)}) \otimes \Phi(x^{(i)})$$

We would like to find eigenfunctions $\hat{\Phi}$ such that

$$(\hat{\Sigma}_n)(\hat{\Phi}) = \lambda \hat{\Phi} \quad (7.2)$$

The question now is, how do we express the above equation in terms of kernels, i.e. how do we ”kernelize” it? Towards this end, we make the following claim:

Claim: Any solution to (7.2) is of the form

$$\hat{\Phi} = \sum_{i=1}^{n} \alpha_i \Phi(x^{(i)})$$

for some weight vector $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. 

7-5
Proof: First, we observe that any solution to (7.2) lies in \( \text{Range}(\hat{\Sigma}_n) \). Linearity, and the nature of \( \Phi(x^{(i)}) \otimes \Phi(x^{(i)}) \) tell us that

\[
\hat{\Sigma}_n(\Phi) = \frac{1}{n} \sum_{i=1}^{n} \langle \Phi(x^{(i)}), \hat{\Phi}(x^{(i)}) \rangle \Phi(x^{(i)})
\]

Therefore, equation (7.2) is equivalent to the following system of equations in \( \alpha \in \mathbb{R}^n \):

\[
\hat{\Sigma}_n \left( \sum_{i=1}^{n} \alpha_i \Phi(x^{(i)}) \right) = \lambda \sum_{i=1}^{n} \alpha_i \Phi(x^{(i)})
\]

For the above set of equations, we have

\[
\text{LHS} = \frac{1}{n} \sum_{i,j=1}^{n} \alpha_i \langle \Phi(x^{(j)}), \Phi(x^{(i)}) \rangle \Phi(x^{(j)})
\]

Using the fact that \( \langle \Phi(x^{(j)}), \Phi(x^{(i)}) \rangle = \langle \Phi(x^{(i)}), \Phi(x^{(j)}) \rangle = \mathbb{K}(x^{(i)}, x^{(j)}) \), the above system of equations may be written as

\[
\frac{1}{n} \sum_{i,j=1}^{n} \alpha_i \mathbb{K}(x^{(i)}, x^{(j)}) \Phi(x^{(j)}) = \lambda \sum_{i=1}^{n} \alpha_i \Phi(x^{(i)})
\]

Taking inner products with \( \Phi(x^{(l)}) \), \( l = 1, \ldots, n \), we get

\[
\frac{1}{n} \sum_{i,j=1}^{n} \alpha_i \mathbb{K}(x^{(i)}, x^{(j)}) \mathbb{K}(x^{(j)}, x^{(l)}) = \lambda \sum_{i=1}^{n} \alpha_i \mathbb{K}(x^{(i)}, x^{(l)}).
\]

We now have a set of \( n \) linear equations in the vector \( \alpha \in \mathbb{R}^n \). In matrix-vector form, it can be written very simply as

\[
K^2 \alpha = \lambda n K \alpha,
\]

where \( K \in \mathbb{R}^{n \times n} \) is the familiar kernel Gram matrix, with entries \( K_{ij} = \mathbb{K}(x^{(i)}, x^{(j)}) \).

The only solutions of this equation that are of interest to us are those that satisfy

\[
K \alpha = \lambda n \alpha.
\]

This is simply an eigenvalue/eigenvector problem in the matrix \( K \). \( \square \)