CR12: Statistical Learning & Applications

Examples of Kernels and Unsupervised Learning

Lecturer: Julien Mairal *Scribes: Rémi De Joannis de Verclos & Karthik Srikanta*

1 Kernel Inventory

Linear Kernel

The linear kernel is defined as $K(x,y) = x^{\mathrm{T}}y$, where $K : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$. Let us show that the corresponding reproducible kernel Hilbert space is $\mathcal{H} = \{y \mapsto w^{\mathrm{T}}y \mid w \in \mathbb{R}^p\}$ where f_w is the function which maps y to $w^{\mathrm{T}}y$, with the inner product $< f_{w_1}, f_{w_2} >_{\mathcal{H}} = w_1^{\mathrm{T}}w_2$. We check the following conditions:

(1) $\forall x \in \mathbb{R}^p$,

$$\{K_x \mid y \mapsto K(x,y)\} \in \mathcal{H}$$

(2) $\forall f_w \in \mathcal{H}$, $\forall x \in \mathbb{R}^p$,

$$f_w(x) =< f_w, K_x >_{\mathcal{H}} = w^{\mathrm{T}}x$$

(3) K is positive semidefinite i.e. $\forall m \in \mathbb{R}$, $\forall (x_1, \ldots, x_m) \in (\mathbb{R}^p)^m$,

$$[K]_{i,j} = K(x_i, x_j) \text{ and } \forall \alpha \in \mathbb{R}^m, \alpha^{\mathrm{T}}K\alpha \geq 0$$

Polynomial kernel with degree 2

The polynomial kernel is given by $K(x,y) = (x^{\mathrm{T}}y)^2$. We now see that it is positive semidefinite as,

$$\begin{aligned}
\sum_{i,j} \alpha_i\alpha_j K(x_i,x_j)^2 &= \sum_{i,j} \alpha_i\alpha_j \mathrm{Trace}(x_i x_i^{\mathrm{T}} x_j x_j^{\mathrm{T}}) \\
&= \sum_{i,j} \alpha_i\alpha_j < x_i x_i^{\mathrm{T}}, x_j x_j^{\mathrm{T}} >_F \\
&= \sum_{i,j} < \alpha_i x_i x_i^{\mathrm{T}}, \alpha_j x_j x_j^{\mathrm{T}} >_F \\
&= \mathrm{Trace}\left(\left(\sum_i \alpha_i x_i x_i^{\mathrm{T}}\right)\left(\sum_j \alpha_j x_j x_j^{\mathrm{T}}\right)\right) \\
&= \mathrm{Trace}\left(Z^T Z\right) \quad \left(\text{ where } Z = \left(\sum_i \alpha_i x_i x_i^{\mathrm{T}}\right)\right) \\
&= || \, Z \, ||_F^2 \\
&\geq 0
\end{aligned}$$

It is important here to define the Frobenius norm. $|| Z ||_F = \sqrt{\sum_{i,j} Z_{i,j}^2}$. Also note that $< Z_1, Z_2 >= \text{Trace}(Z_1^{\mathrm{T}} Z_2)$.

Now we would like to find $\mathcal{H}$ such that (1) and (2) mentioned above are satisfied. To satisfy (1), we need $f_x : y \mapsto K(x,y) \in \mathcal{H}, \forall x \in \mathbb{R}^p$ and,

$$\begin{aligned} f_x(y) &= (x^{\mathrm{T}} y)^2 \\ &=< xx^{\mathrm{T}}, yy^{\mathrm{T}} >_F \\ &= y^{\mathrm{T}}(xx^{\mathrm{T}})y \end{aligned}$$

Because $\mathcal{H}$ is a Hilbert space, we also have $\forall \beta$ in $\mathbb{R}^n$, $\forall (x_1, \ldots, x_n) \in (\mathbb{R}^p)^n$, $y \mapsto y^{\mathrm{T}} \left(\sum_{i=1}^n \beta_i x_i x_i^{\mathrm{T}} \right) y$ is in $\mathcal{H}$. A good candidate for $\mathcal{H}$ is thus $\mathcal{H} : \{ y \mapsto y^{\mathrm{T}} A y, A \text{ symmetric } \in \mathbb{R}^{p \times p} \}$ with $< f_A, f_B >_{\mathcal{H}} = \text{Trace}(A^{\top} B)$. This is easy to check: $f_A(x) = x^{\mathrm{T}} A x =< A, xx^{\mathrm{T}} >_F$, $K_x = f_{xx^{\mathrm{T}}}$, which implies $f_A(x) =< f_A, K_x >$.

Gaussian kernel

The Gaussian kernel is given by $K(x,y) = e^{-\frac{\|y-x\|^2}{\omega^2}} = k(x-y)$.
The corresponding r.k.h.s is more involved. We will admit that if we define $< f, g >_{\mathcal{H}} = \frac{1}{(2\pi)^p} \int \frac{\hat{f}(w)\hat{g}(w)^*}{\hat{k}(w)} dw$, we have,

$$\mathcal{H} = \{ f \mid f \text{ is integrable, continuous and } < f, f >_{\mathcal{H}} < +\infty \}$$

The min kernel

Let $K(x,y) = \min(x,y)$. $\forall x, y \in [0,1]$.
Let is show that K is positive semidefinite: $\forall \alpha \in \mathbb{R}^n, \forall (x_1, .., x_n) \in [0,1]^n$ we have,

$$\begin{aligned} \sum_{i,j} \alpha_i \alpha_j \min(x_i, x_j) &= \sum_{i,j} \alpha_i \alpha_j \int_0^1 \mathbb{1}_{t \le x}(t) \mathbb{1}_{t \le y}(t) \\ &= \int_0^1 \left(\sum_{i=1}^n \alpha_i \mathbb{1}_{t \le x_i} \right) \left(\sum_{j=1}^n \alpha_j \mathbb{1}_{t \le x_j} \right) dt \\ &= \int_0^1 Z(t)^2 dt \quad \left(\text{where } Z(t) = \left(\sum_{j=1}^n \alpha_j \mathbb{1}_{t \le x_j} \right) = \left(\sum_{i=1}^n \alpha_i \mathbb{1}_{t \le x_i} \right) \right) \\ &\ge 0 \end{aligned}$$

It is thus appealing to believe that $\mathcal{H} = \{ f \mid f \in L^2[0,1] \}$ and $< f, g >= \int_0^1 f(t) g(t) dt$. However, this is a mistake since $< f, f >= 0$ does not imply that $f = 0$. In fact,

$$\mathcal{H} = \{ f : [0,1] \to \mathbb{R} \mid \text{ continuous and differentiable almost everywhere and } f(0) = 0 \}$$

Now we have $< f, g >_{\mathcal{H}} = \int_0^1 f'(t) g'(t) dt$ and thus $< f, f >_{\mathcal{H}} = 0 \Leftrightarrow f' = 0 \Leftrightarrow f = 0$.

We can now check the remaining conditions

- $K_x : y \mapsto \min(x,y) \in \mathcal{H}$

- $f \in \mathcal{H}, x \in [0;1]$

$$f(x) =< f, K_x >= \int_0^1 f'(t)\mathbb{1}_{t<=x}dt = \int_0^1 f'(t)dt = f(x) - f(0) = f(x)$$

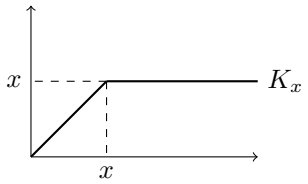


Figure 1: Graph of K_x

Histogram kernel

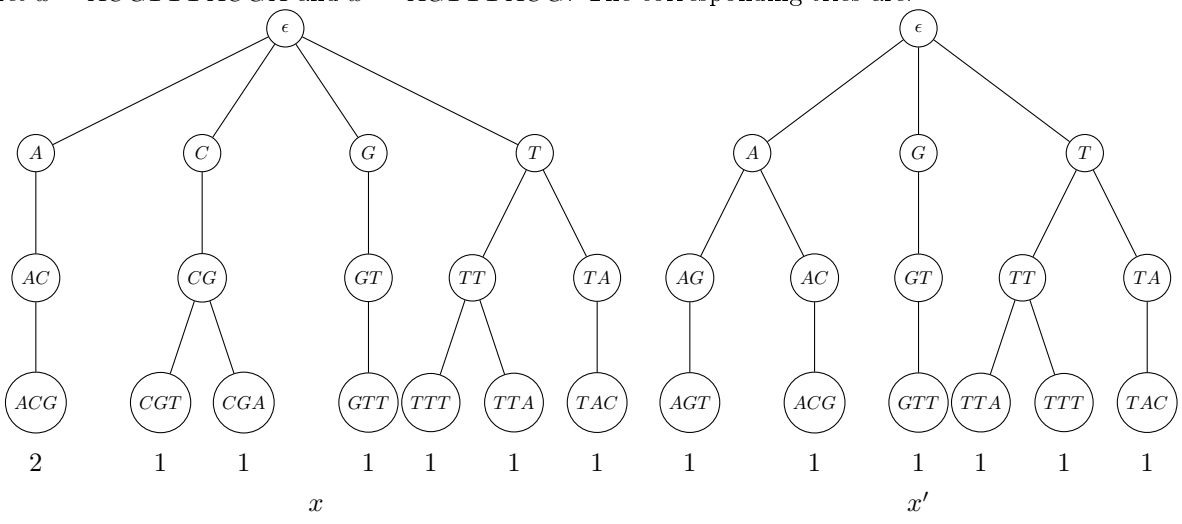
The Histogram kernel is given by $K(h_1, h_2) = \sum_{j=1}^{k} \min(h_1(j), h_2(j))$ where $h_1 \in [0,1]^k$. It is notably used in computer vision for comparing histograms of visual words.

Spectrum kernel

We now try to motivate the spectrum kernel through biology. There are 25000 human protein–coding genes. In other words there are $3,000,000,000$ bases of building blocks A, T, C, G. Consider a sequence $u \in \mathcal{A}^k$ of size k (where in this case we assume $A = \{A, T, C, G\}$). Let $\phi_u(x)$ = number of occurrences of u in x. We define the spectrum kernel as $K(x, x') = \sum_{u \in \mathcal{A}^k} \phi_u(x)\phi_u(x') =< \phi(x), \phi(x') >$ where $\phi(x) \in \mathbb{R}^{|A|^k}$.

We are interested in computing $K(x, x')$ efficiently.
$K(x, x') = \sum_{i=1}^{|x|-k+1} \sum_{j=1}^{|x'|-k+1} \mathbb{1}_{x[i,i+k-1]=x[j,j+k-1]}$. The computation of this can be done in $\mathcal{O}(\mid x \mid\mid x' \mid k)$ by using a trie. We demonstrate this with an example.

Let $x = ACGTTTACGA$ and $x' = AGTTTACG$. The corresponding tries are:



Algorithm 1 Process Node(v)

Input: Trie representation for sequences x and x'.
Output: Compute $K(x, x')$.

```
if v is a leaf then
    K(x, x') = K(x, x') + i(v_x) · i(v_{x'})
else
    for All children in common of v_x, v'_x do
        Process Node(children).
    end for
end if
Repeat Process for Node(r).
```

Exercises

1. Show that if K_1 and K_2 are positive semidefinite then,
   (a) $K_1 + K_2$ is positive semidefinite.
   (b) $\alpha K_1 + \beta K_2$ is positive semidefinite, where $\alpha, \beta > 0$.
2. Show that if $(K_n)_{n>0} \to K$ (pointwise), then K is positive semidefinite.
3. Show that $K(x, y) = \frac{1}{1-\min(x,y)}$ is positive semidefinite for all $x, y \in [0, 1[$.

Walk kernel

We define $G' = (V', E')$ is a subgraph of G if $V' \subseteq V$, $E' \subseteq E$ and labels(V) = labels(V'). Let $\phi_U(G) = \{$ Number of times U appears as a subgraph of $G\}$. The subgraph kernel is defined as,

$$K(G, G') = \sum_U \phi_U(G)\phi_U(G')$$

Unfortunately computing this in NP-complete.

We now define the walk kernel. We recall some definitions first. A walk is an alternating sequence of vertices and connecting edges. Less formally a walk is any route through a graph from vertex to vertex along edges. A walk can end on the same vertex on which it began or on a different vertex. A walk can travel over any edge and any vertex any number of times. A path is a walk that does not include any vertex twice, except that its first vertex might be the same as its last. Now let,
$W_k = \{$walks in G of size $k\}$ $S_k = \{$sequence of labels of size $k\}$ $=\mid \mathcal{A} \mid^k$ So $\forall s \in S_k, \phi_s(G) = \sum_{w \in W_k(G)} \mathbb{1}_{\text{labels}(w)=s}$.
This gives us $K(G_1, G_2) = \sum_{s \in S} \phi_s(G_1)\phi_s(G_2)$.

Computing the walk kernel seems hard at first sight, but it can be done efficiently by using the product graph, defined as follows. Given graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, we define,

$$G_1 \times G_2 = (\{(v_1, v_2) \in V_1 \times V_2 \text{ with same labels}\}\{((v_1, v_2), (v'_1, v'_2)) \text{ where } ((v_1, v'_1) \in E_1 \text{ and } (v_2, v'_2) \in E_2)\})$$

There is a bijection between walks in $G_1 \times G_2$ and walks in G_1 and G_2 with same labels. More formally,

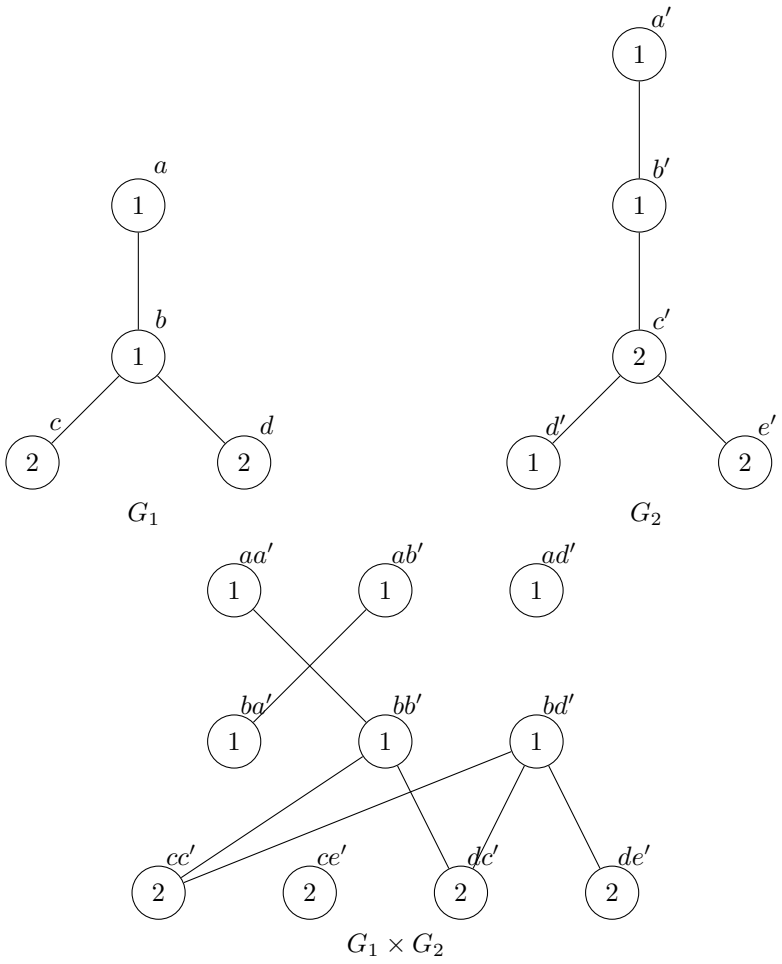


Figure 2: Example of graph product

$$\begin{aligned} K(G_1, G_2) &= \sum_{w_1 \in W_k(G_1)} \sum_{w_2 \in W_k(G_2)} \mathbb{1}_{l(w1)=l(w_2)} \\ &= \sum_{w \in W_k(G_1 \times G_2)} 1 \end{aligned}$$

Exercise Let A be the adjacency matrix of $G \in \mathbb{R}^{|V| \times |V|}$,

1. Prove that number of walks of size k starting at i and ending at j is $[A^k]_{i,j}$.

2. Show that $K(G_1, G_2)$ can be computed in polynomial time.

2 Unsupervised Learning

Unsupervised learning consists of discovering the underlying structure of data without labels. It is useful for many tasks, such as removing noise from data (preprocessing), interpreting and visualizing the data, compression...

Principal component analysis

We have a centered data set $X = [x_1, \dots, x_n] \in \mathbb{R}^{p \times n}$.
Consider the projection of x onto the direction $w \in \mathbb{R}^p$, then $x \mapsto \frac{w^{\mathrm{T}}x}{\|w\|_2^2} \cdot \frac{w}{\|w\|_2^2}$. We observe that the empirical variance captured by w is $\mathrm{Var}'(w) = \frac{1}{n}\sum_{i=1}^n \frac{(w^{\mathrm{T}}x_i)^2}{\|w\|_2^2}$.

We provide below the PCA (Principle Component Analysis) algorithm:

Algorithm 2 PCA

1: Suppose we have $(w_1, \dots, w_{i-1})$
2: Construct $w_i \in \arg\max Var(w)$ under the condition $w_i \perp (w_1, \dots, w_{i-1})$

Lemma: $\frac{w_i}{\|w_i\|_2}$ are successive vectors of XX^{T} ordered by decreasing eigenvalue.
Proof: We have $w_1 = \underset{\|w\|_2=1}{\arg\max} \frac{1}{n}\sum_{i=1}^n w^{\mathrm{T}}x_i x_i^{\mathrm{T}} w = \underset{\|w\|_2=1}{\arg\max} w^{\mathrm{T}} X X^{\mathrm{T}} w$.
XX^{T} is symmetric, so it can be diagonalized into $XX^{\mathrm{T}} = USU^{\mathrm{T}}$ where U is orthogonal and S is diagonal.

$$\begin{aligned} w_1 &= \underset{\|w\|_2=1}{\arg\max}\ w^{\mathrm{T}} U S U^{\mathrm{T}} w \\ &= \underset{\|w\|_2=1}{\arg\max}\ (U^{\mathrm{T}}w)^{\mathrm{T}} S (U^{\mathrm{T}}w) \\ &= \left(\underset{\|z\|_2=1}{\arg\max}\ z^{\mathrm{T}} S z \right) \text{ where } z = w^{\mathrm{T}}U \end{aligned}$$

For w_i, $i > 1$ we have,

$$\begin{aligned} w_i &= \underset{\|w\|_2=1, w \perp u_1, \dots, u_{i-1}}{\arg\max}\ w^{\mathrm{T}} U S U^{\mathrm{T}} w \\ &= \underset{\|w\|_2=1, w \perp u_1, \dots, u_{i-1}}{\arg\max}\ (U^{\mathrm{T}}w)^{\mathrm{T}} S (U^{\mathrm{T}}w) \\ &= \underset{\|z\|_2=1, z=e_1, \dots, e_{i-1}}{\arg\max}\ z^{\mathrm{T}} S z \quad \blacksquare \end{aligned}$$

As a consequence we have PCA can be computed with SVD.

Theorem(Aqart - Young theorem)
PCA (SVD) provides the best low–rank approximation.

$$\min_{X' \in \mathbb{R}^{n \times p}} \|X - X'\|^2, rank(X') < k$$

Proof: Any matrix X' of rank k can be written $X' = \sum_{i=1}^k s_i u_i v_i^{\mathrm{T}} = U_k S_k V_k^{\mathrm{T}}$ where $U_k \in \mathbb{R}^{p \times k}$, $V_k \in \mathbb{R}^{k \times n}$,

$U_k V_k = I$ and $S_k \in M_k(R)$ is diagonal. Thus,

$$\min_{\substack{X \in \mathbb{R}^{n\times p} \\ rank(X')<k}} \|X - X'\|^2 = \min_{\substack{U^{\mathrm{T}}U=I \\ V^{\mathrm{T}}V=I \\ S \text{ diagonal} \\ rank(USV^{\mathrm{T}})=k}} \|X - USV^{\mathrm{T}}\|_p^2 \quad \blacksquare$$

Lemma: Let $U \in \mathbb{R}^{p\times m}$ and $V \in \mathbb{R}^{n\times m}$ where $m = \min(n, p)$. If $U^{\mathrm{T}}U = I$ and $V^{\mathrm{T}}V = I$ then $Z_{ij} = U_i V_j^{\mathrm{T}}$ form an orthogonal basis of $\mathbb{R}^{p\times n}$ for the Frobenius norm.
Proof: It suffices to check the scalar product:

$$\begin{aligned} < Z_{ij}, Z_{kl} > &= \mathrm{Trace}(Z_{ij}^{\mathrm{T}} Z_{kl}) \\ &= \mathrm{Trace}(V_j U_i^{\mathrm{T}} U_k V_l^{\mathrm{T}}) \\ &= (V_j^{\mathrm{T}} V_l)(U_j^{\mathrm{T}} U_k) \\ &= \mathbb{1}_{\{j=k \text{ and } i=l\}} \quad \text{since } U \text{ and } V \text{ are orthogonal.} \blacksquare \end{aligned}$$

We have $X = \sum_{i=1}^m s_i Z_{ii}$ and because of the lemma we can write X' in the basis Z_{ij}:

$$X' = \sum_{i=1}^m s'_{ij} Z_{ij}$$

Under the condition $rank(X') = r$,
$\|X - X'\|^2$ is minimal iff $\sum_{i,j}(s'_{ij} - s_{ij})^2$ is minimal
iff $\sum_{i=1}^n (s'_{ii} - s_i)^2 + \sum_{i\neq j}(s'_{ij})^2$ is minimal
Then, the first term should be minimized by taking $s'_{ii} = s_i$ for the k longest values of s_i with respect to the rank constaint and the second term should be set to 0.

Kernel PCA

If we have the kernel $\phi : x \mapsto \phi(x)$
$Var(f) = \frac{1}{n}\sum_{i=1}^n \frac{f^2(x_i)}{\|f\|_H^2}$ assuming $\phi(x_i)$ are centered in the feature space

$$f_i \in \underset{\substack{f_1 \perp f_2, \ldots, f_{i-1} \\ f_i \in H}}{\arg\max} \; Var(f) \tag{1}$$

There are few remaining questions

1. What does it mean to have $\phi(x_i)$ centered?
2. How do we solve (1)

1. Having $\phi(x_i)$ centered means that you want to implicitely replace $\phi(x_i)$ by $\phi(x_i) - m$ where $m = \frac{1}{n}\sum_{i=1}^n \phi(x_i)$ is the average of ϕ. If our kernel is $K(x_i, x_j) =< \phi(x_i), \phi(x_j) >_H$, the new kernel becomes:

$$\begin{aligned} K_c(x_i, x_j) &=< \phi(x_i) - m, \phi(x_j) - m > \\ &=< \phi(x_i), \phi(x_j) > - < \frac{1}{n}\sum_{l=1}^n \phi(x_l), \phi(x_i) > - < \frac{1}{n}\sum_{l=1}^m \phi(x_l), \phi(x_j) > + \frac{1}{n^2}\sum_{i,k} < \phi(x_l), \phi(x_k) > \end{aligned}$$

In term of matrices, it is writen $K_c = (I - \frac{1}{n}\mathbf{1})K(I - \frac{1}{n}\mathbf{1})$ where $\mathbf{1}$ is the square matrix with all coefficients equal to 1.

2. Due to the representation theorem, any solution of (1) has the form $f : x \mapsto \sum_{i=1}^{n} \alpha_i K(x_i, x)$. What we need are the α's. We have $\|f\|_H^2 = \alpha^{\mathrm{T}} K\alpha$, $f(x_i) = [K\alpha]_i$ and $< f, f_l >= 0 = \alpha^{\mathrm{T}} K\alpha_l$

We want to find $\max_{\alpha \in R^n} \frac{1}{n} \frac{\|K\alpha\|_2^2}{\alpha^{\mathrm{T}} K\alpha}$ s.t. $\alpha^{\mathrm{T}} K\alpha_l = 0, \forall l < i$. i.e. $\max_\alpha \frac{\alpha^{\mathrm{T}} K^2 \alpha}{\alpha^{\mathrm{T}} K\alpha}$ and $\alpha^{\mathrm{T}} K\alpha_l = 0, \forall l < i$. K is symetric so $K = US^2U^{\mathrm{T}}$ where U is orthogonal and S diagonal. We want $\max_\alpha \frac{\alpha^{\mathrm{T}} US^4U^{\mathrm{T}}\alpha}{\alpha^{\mathrm{T}} US^2U^{\mathrm{T}}\alpha} = \max_{\beta \in \mathbb{R}^n} \frac{\beta^{\mathrm{T}} S^2 \beta}{\|\beta\|_2^2}$ where $\beta = SU^{\mathrm{T}}\alpha$, $\beta_l = SU^{\mathrm{T}}\alpha_l$. If we assume that the eigenvalues are ordered in S, the solutions are the duals of $\beta_l = e_l$.

Recipe:

1. Center the kernel matrix.

2. $K_L = US^2U^T$.

3. $a_l = \frac{1}{D^2} U_i$ (eigenvalue decomposition)

4. $f_i(x_i) = [K_i\alpha]_i$

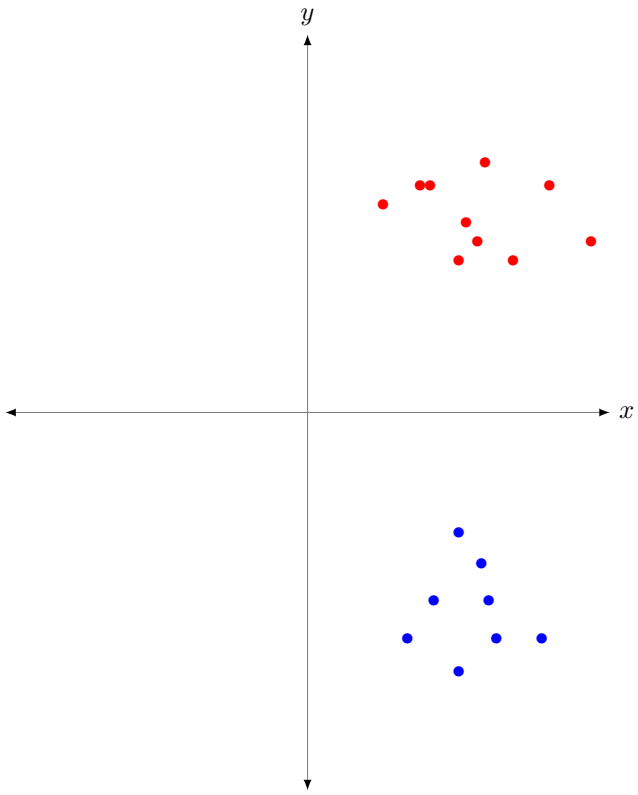
K-means clustering

Let $X = [X_1, .., X_n] \in \mathbb{R}^{p \times n}$ be data points and we would like to form clusters $C = [C_1, ..., C_k] \in \mathbb{R}^{p \times k}$ where each of these are centroids of the cluster.

Our goals are to:

1. Learn C.

2. Assign each data point to a centroid C_j.



A popular way to do this is by K-means algorithm:

Algorithm 3 K-means

1: **for** $i = 1, \ldots, n$ **do**
2: Assign $l_i \leftarrow \text{argmin}_{j=1,..,K} \|X_i - C_j\|_2^2$
3: **end for**
4: **for** $j = 1, \ldots, k$ **do**
5: Re–estimate centroids by $C_j \leftarrow \frac{1}{n_j} \sum_{i \text{ such that } l_i = j} X_i$
6: **end for**

The interpretation of the algorithm can be seen as following:
We are trying to find C such that we have $\min_{C,l_i} \sum_{i=1}^{n} \|X_i - C_{l_i}\|_2^2$.
Now given some fixed labels,

$$\nabla_{c_1} f(C_1) = \sum_{i \text{ such that } l_i = 1} 2[C_1 - X_i] = 0$$

. Note that K–means is not an optimal algorithm because $\min_{C,l_i} \sum_{i=1}^{n} \|X_i - C_{l_i}\|_2^2$ is a non-convex optimization problem with several local minima.

Kernel K-means

Let us define the Kernel for this to be $\min_{c,l \in \mathcal{H}} \sum_{i=1}^{n} \|\phi(X_i) - C_{l_i}\|_{\mathcal{H}}^2$.
Given some labels (re–estimation)
$C_j \leftarrow \frac{1}{n_j} \sum_{i \text{ such that } l_i = j} \phi(X_i)$

Now $l_i \leftarrow \arg\min_j \|\phi(X_i) - \frac{1}{n_j}\sum_{l\in S_j}\phi(X_l)\|^2 = \|\phi(x_i)\|^2 + \|C_j\|^2 - \frac{2}{n_j}\sum_{i\in S_j}K(x_i, x_j)$.

Mixture of Gaussians

The Gaussian distribution is given by:

$$N(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{p/2} \mid \Sigma \mid^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)\Sigma^{-1}(x-\mu)\right)$$

Assume that the data is generated by the following procedure:

1. Random class assignment: C: $\mathbb{P}(C = C_j) = \pi_j$ where $\sum_{j=1}^k \pi_j = 1, \pi_j \geq 0$.
2. $x \approx N(x, \mu_i, \sigma_i)$

The parameters are:

- $\pi_1, \ldots, \pi_k$
- $(\mu_1, \sigma_1), \ldots, (\mu_i, \sigma_i)$

If we observe $X_1, \ldots, X_n$, can we infer the parameters (π, μ, Σ) ?
Let θ represent the parameters.
Let $\hat{\theta} = \arg\max_\theta \mathbb{P}_\theta(X_1, \ldots, X_n)$ (Maximum $= \arg\max_\theta \prod_{i=1}^n \mathbb{P}_\theta(X_i)$

One algorithm to get an approximate solution is called EM for "Expectation-maximization", which iteratively increases the likelihood

Algorithm 4 Expectation Maximizer

1: Define $L_\theta = -\log \mathbb{P}_\theta(X_1, ..., X_n) = \sum_{i=1}^n -\log \mid \mathbb{P}_\theta(X_i) \mid$.
2: **for** $1 = 1, \ldots, n$ **do**
3: E-step: find q given θ fixed (auxiliary soft assignment)
4: M-step: Maximize some function $l(\theta, q)$ with q fixed.
5: **end for**
6: Repeat step 2 until convergence.

Explanations:
At E-step $q_{i,j} = \frac{\prod_j N(x_i|\mu_j,\sigma_j)}{\sum N(x_i|\mu_{j'};\sigma_{j'})}$. Notice that $\forall i, \sum_{j=1}^k q_{ij} = 1$. This can be interpreted as a soft assignment of very data point to the classes
At M-step we have:

$$\pi_j \leftarrow \frac{1}{n}\sum_{i=1}^n q_{ij}$$

$$\mu_j \leftarrow \sum_{i=1}^n q_{ij}x_i$$

$$\Sigma_j \leftarrow \sum_{i=1}^m q_{ij}(x_i - \mu_j(x_i - \mu_j)^{\mathrm{T}}$$

The key ingredient here is the use of Jensen inequality, but we will not have the time to present the details in this course.