This homework is an exan that was given during a previous iteration of this course. I have voluntarily kept the presentation as it was for the exam.

1 Multivariate Fisher Kernel

The Fisher kernel is a kernel for data in $\mathbb{R}^d$ which is based on a probabilistic model. Specifically, let $x \in \mathbb{R}^d$, and let p_θ be a probability density function indexed by a parameter vector $\theta \in \mathbb{R}^k$. First, fix some $\theta_0 \in \mathbb{R}^k$ and define the Fisher score of a data point x, by $\phi(x)$, the gradient of the logarithm of p_θ evaluated at $\theta = \theta_0$. That is,

$$\phi(x) = \nabla_\theta \ln p_\theta(x)|_{\theta=\theta_0} \tag{1}$$

Thus $\phi : \mathbb{R}^d \to \mathbb{R}^k$.

http://yongyuan.name/blog/fim-fisher-kernel.html

https://wiseodd.github.io/techblog/2018/03/11/fisher-information/

Next, define the Fisher information matrix as the following expected value:

$$I = E_{p_{\theta_0}}[\phi(X)\phi^T(X)] \tag{2}$$

Note that I is a (k, k) matrix. You can use the fact that this matrix is symmetric and positive definite. Finally, define the Fisher kernel as

$$K(x, y) = \phi(x)^T I^{-1} \phi(y) \tag{3}$$

for any couple of points (x, y), both in $\mathbb{R}^d$

1. Verify that K is symmetric

$$\begin{aligned} E_{p_{\theta_0}}[\phi(X)] = E_{p_{\theta_0}}\left[\nabla_\theta \ln p_\theta(x)|_{\theta=\theta_0}\right] &= \int \nabla \ln p_\theta(x)\, p_\theta(x)\, \mathrm{d}x = \int \frac{\nabla p_\theta(x)}{p_\theta(x)} p_\theta(x)\, \mathrm{d}x \\ &= \int \nabla p_\theta(x)\, \mathrm{d}x = \nabla \int p_\theta(x)\, \mathrm{d}x = \nabla 1 = 0 \end{aligned}$$

$I = E_{p_{\theta_0}}[\phi(X)\phi^T(X)] = E[\phi(X)^2] = V[\phi(X)^2] - E[\phi(X)]^2 = V[\phi(X)^2] - 0 \geq 0$

For $\underset{(k,k)}{I}$ is symmetric and positive definite, $I^{-1} = L^T L$ by Cholesky decomposition, where L is a lower triangular matrix with real and positive diagonal entries.

$$\begin{aligned} K(x, y) = \phi(x)^T I^{-1} \phi(y) = \phi(x)^T L^T L \phi(y) &= (L\phi(x))^T (L\phi(y)) \\ &= \langle L\phi(x), L\phi(y)\rangle_{\mathcal{H}} = \langle L\phi(y), L\phi(x)\rangle_{\mathcal{H}} = K(y, x) \end{aligned}$$

where $\mathcal{H}$ is a Hilber Space. $L\phi(\cdot)$ is a function $\mathcal{X} \to \mathbb{R}$

2. Verify that K is positive definite

Choose $x_1, ..x_n \in \mathcal{X}$; $\alpha_1, ..\alpha_n \in \mathbb{R}$

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) = \sum_{i,j=1}^{n} \alpha_i \alpha_j (L\phi(x_i))^T (L\phi(x_j)) = (\sum_{i=1}^{n} \alpha_i L\phi(x_i))^T (\sum_{j=1}^{n} \alpha_j L\phi(x_j)) = \|\alpha L\phi(x)\|_{\mathcal{H}}^2 \geq 0$$

3. Consider the following multivariate Normal model with given invertible covariance matrix Λ^{-1}: $p_\theta(x) = (2\pi)^{-d/2}(\det\Lambda)^{1/2}\exp\left(-\frac{1}{2}(x-\theta)^T\Lambda(x-\theta)\right)$ Show that $\phi(x) = \Lambda(x-\theta_0)$

$$\begin{aligned}\phi(x) &= \nabla_\theta \ln p_\theta(x)|_{\theta=\theta_0} = \nabla_\theta \ln(2\pi)^{-d/2}(\det\Lambda)^{1/2}\exp\left(-\frac{1}{2}(x-\theta_0)^T\Lambda(x-\theta_0)\right)\\ &= \nabla_\theta - \frac{d}{2}\ln(2\pi) + \frac{1}{2}(\det\Lambda) - \frac{1}{2}(x-\theta_0)^T\Lambda(x-\theta_0)\\ &= -\frac{2}{2}\Lambda(x-\theta_0)(-1) = \Lambda(x-\theta_0)\end{aligned}$$

4. Compute the Fisher information matrix I for this model

$$\begin{aligned}I &= E_{p_{\theta_0}}[\phi(X)\phi^T(X)] = E_{p_\theta}[(\Lambda(x-\theta))^T\Lambda(x-\theta)]\\ &= \Lambda E_{p_\theta}[(x-\theta)^T(x-\theta)]\Lambda = \Lambda^3\end{aligned}$$

5. Compute the Fisher kernel for this model

$$\begin{aligned}K(x,y) &= \phi(x)^T I^{-1}\phi(y) = (\Lambda(x-\theta_0))^T(\Lambda^3)^{-1}\Lambda(x-\theta_0)\\ &= (x-\theta_0)^T\Lambda(\Lambda\Lambda\Lambda)^{-1}\Lambda(y-\theta_0) = (x-\theta_0)^T\Lambda^{-1}(y-\theta_0)\end{aligned}$$

2 Optimal ordering

We consider a binary classification problem where the training set is $\mathcal{D} = \{(x_i, y_i), 1 \le i \le n\}$, with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$ is the class. For convenience, we consider the following subset of indices

$$I_- = \{i, 1 \le i \le n, y_i = -1\} \quad (4)$$
$$I_+ = \{i, 1 \le i \le n, y_i = +1\} \quad (5)$$

Moreover, notate n_- the number of elements in I_-, and n_+ the number of elements in I_+. Thus

$$n_- + n_+ = n \quad (6)$$

Our objective is to use $\mathcal{D}$ to construct a function $f : \mathbb{R}^d \to \mathbb{R}$ that assign larger values to the data points in the positive class than to the data points in the negative class while being smooth in some sense.

Thus, we propose to choose f in a RKHS with a kernel K that minimizes the following functional:

$$J_0(f) = \frac{1}{n_- n_+}\sum_{i\in I_-}\sum_{j\in I_+}\mathbb{I}_{f(x_i)>f(x_j)} + \lambda||f||_H^2 \quad (7)$$

where $\mathbb{I}_u$ is the indicator function of the event u. It takes the value 1 if the event u occurs and 0 otherwise and $\lambda > 0$.

Since J_0 is not convex, we propose instead to minimize

$$J(f) = \frac{1}{n_- n_+}\sum_{i\in I_-}\sum_{j\in I_+}(1 - (f(x_j) - f(x_i))) + \lambda||f||_H^2 \quad (8)$$

1. Show that the minimum is achieved for a function f index by a parameter $\alpha \in \mathbb{R}^d$ and of the form

$$f(x) = \sum_{i=1}^n \alpha_i K(x_i, x) \quad (9)$$

Note: do not invoke the representer theorem. Instead, use the key elements in the proof of this theorem applied to this particular case.

Let $v = \text{span}[k(\cdot, x_i), .., k(\cdot, x_n)]$ $\mathcal{V}$ is closed linear subspace of $\mathcal{H}$. Then all minimizers of J belong to $\mathcal{V}$. Thus, there is an unique decomposition $f = f_v + f_\perp$ with $f_v \in \mathcal{V}$

$\forall f \in \mathcal{V}$, $\langle f_\perp, g\rangle = 0$

$$\|f\|_{\mathcal{H}}^2 = \|f_v + f_\perp\|_{\mathcal{H}}^2 = \langle f_v + f_\perp, f_v + f_\perp\rangle = \langle f_v, f_v\rangle + \langle f_\perp, f_\perp\rangle + \underbrace{2\langle f_v, f_\perp\rangle}_{0} = \|f_v\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2$$

$$f(x_i) = \langle f, k(\cdot, x_i)\rangle = \langle f_v + f_\perp, k(\cdot, x_i)\rangle = \langle f_v, k(\cdot, x_i)\rangle + \underbrace{\langle f_\perp, k(\cdot, x_i)\rangle}_{0} = f_v(x_i)$$

For f is strictly increasing

$$\begin{aligned} J(f) - J(f_v) &= \frac{1}{n_- n_+}\sum_{i\in I_-}\sum_{j\in I_+}[1 - (f(x_j) - f(x_i))] + \lambda\|f\|_{\mathcal{H}}^2 - \frac{1}{n_- n_+}\sum_{i\in I_-}\sum_{j\in I_+}[1 - (f_v(x_j) - f_v(x_i))] - \lambda\|f_v\|_{\mathcal{H}}^2 \\ &= \lambda\|f\|_{\mathcal{H}}^2 - \lambda\|f_v\|_{\mathcal{H}}^2 = \lambda\|f_\perp\|_{\mathcal{H}}^2 \geq 0 \end{aligned}$$

The representer theorem allows us to reduce the optimization problem to a finite dimensional optimization problem.

Let $\alpha = (\alpha_1, .., \alpha_n)^T \in \mathbb{R}^n$ is the solution of $\min J(f)$, the function $f(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$,$f \in \mathcal{H}$ that minimizes $J(f)$

2. Rewrite then $J(f)$ as a functional $J(\alpha)$ using the notation K for the (n, n) matrix $K_{ij} = K(x_i, x_j)$ and K_i for the i^{th} column of K.

$$\|f\|_{\mathcal{H}}^2 = \langle f, f\rangle_{\mathcal{H}} = \langle\sum_{i=1}^n \alpha_i k(\cdot, x_i), \sum_{j=1}^n \alpha_j k(\cdot, x_j)\rangle_{\mathcal{H}} = \sum_{i,j=1}^n \alpha_i\alpha_j \underbrace{\langle k(\cdot, x_i), k(\cdot, x_j)\rangle_{\mathcal{H}}}_{k(x_i, x_j)} = \underset{(1,n)}{\alpha^T}\underset{(n,n)}{K}\underset{(n,1)}{\alpha}$$

$$f(x_i) = \sum_{j=1}^n \alpha_j k(x_i, x_j) = \sum_{j=1}^n \alpha_j [\underset{(n,n)}{K}]_{i,j} = [K\alpha]_i$$

$$J(\sum_{i=1}^n \alpha_i k(x_i, \cdot)) = \frac{1}{n_- n_+}\sum_{i\in I_-}\sum_{j\in I_+}[1 - ([K\alpha]_j - [K\alpha]_i)] + \lambda\alpha^T K\alpha$$

3. Simplify further the expression $J(\alpha)$ using the notations:

$K_- = \frac{1}{n_-}\sum_{i\in I_-} K_i$; $K_+ = \frac{1}{n_+}\sum_{i\in I_+} K_i$;

$$J(\alpha) = 1 - [\frac{1}{n_+}\sum_{j\in I_+}[K\alpha]_j - \frac{1}{n_-}\sum_{i\in I_-}[K\alpha]_i)] + \lambda\alpha^T K\alpha = 1 - [K_+ - K_-]\alpha + \lambda\alpha^T K\alpha$$

4. Compute $\nabla_\alpha J(\alpha)$, the gradient in α of $J(\alpha)$. Solve for α, assuming that K is invertible.

p.d. K, X is symmetric, $K = K^T$, $X = X^T$; $K = P\Lambda P^T$; $I = PP^T$; Λ is diagonal matrix with $\gamma_1, .., \gamma_n$.

$\lambda > 0$,$\gamma_i > 0$, $K + \lambda I = P(\Lambda + \lambda I)P^T$ is inversible.

$$\nabla_\alpha J = \frac{\partial}{\partial \alpha}(1 - [K_+ - K_-]\alpha + \lambda \alpha^T K \alpha) = -[K_+ - K_-] + \lambda \frac{\partial}{\partial \alpha}\langle \alpha, K\alpha \rangle$$
$$= -[K_+ - K_-] + \lambda(IK\alpha + K^T\alpha) = -[K_+ - K_-] + 2\lambda K\alpha \stackrel{set}{=} 0$$

$$\alpha^\star = (2\lambda K)^{-1}[K_+ - K_-]$$

5. Bonus question: Compute $f^*(x)$, where f^* is the minimizer of $J(f)$ for the linear kernel $K(x, y) = x^T y$. Use the notations $x_- = \frac{1}{n_-}\sum_{i \in I_-} x_i$; $x_+ = \frac{1}{n_+}\sum_{i \in I_+} x_i$

$$\frac{1}{n_+}\sum_{j \in I_+} f(x_l) = \frac{1}{n_+}\sum_{j \in I_+}\sum_{l=1}^{n} \alpha_l k(x_l, x_j) = \sum_{l=1}^{n} \alpha_l x_l \frac{1}{n_+}\sum_{j \in I_+} x_j = \alpha^T K x_+ = f(.)x_+$$

$$\frac{1}{n_-}\sum_{i \in I_-} f(x_l) = \frac{1}{n_-}\sum_{i \in I_-}\sum_{l=1}^{n} \alpha_l k(x_l, x_i) = \sum_{l=1}^{n} \alpha_l x_l \frac{1}{n_-}\sum_{i \in I_-} x_j = \alpha^T K x_- = f(.)x_-$$

$$J(f) = 1 - (f(.)x_+ - f(.)x_-) + \lambda f(.)^T f(.)$$

$$\nabla_{f(.)} J = \frac{\partial}{\partial f(.)}[1 - (f(.)x_+ - f(.)x_- + \lambda f(.)^T f(.)] = -[x_+^T - x_-^T] + \lambda \frac{\partial}{\partial \alpha}\langle f(.), f(.)\rangle$$
$$= -[x_+^T - x_-^T] + 2\lambda f(.) \stackrel{set}{=} 0$$

$$f^\star(x) = (2\lambda)^{-1}[x_+^T - x_-^T]$$