Homework exercises Advanced Learning Models 2016-2017

Jakob Verbeek & Julien Mairal
INRIA Grenoble – Université Grenoble-Alpes

January 11, 2017

Exercise 1: Fisher kernel for univariate Gaussian density

Suppose a univariate Gaussian density model $p(x) = \mathcal{N}(x; \mu, \sigma^2)$.

1. Compute the partial derivatives $\frac{\partial \ln p(x)}{\partial \mu}$ and $\frac{\partial \ln p(x)}{\partial \sigma}$.

Let $g(x)$ be the two dimensional gradient vector that concatenates the two partial derivatives.

2. Compute the Fisher information matrix $F = \int_x p(x) g(x) g(x)^\top$.

3. Show that $\int_x p(x) g(x) = 0$.

4. Compute the Fisher vector $h = F^{-\frac{1}{2}} g$.

Exercise 2: Fisher kernel for univariate Gaussian mixture density

Suppose a univariate Gaussian mixture density model $p(x) = \sum_{i=1}^K w_i \mathcal{N}(x; \mu_i, \sigma_i^2)$. Where the mixing weights are parameterized as $w_i = \exp(\alpha_i)/\sum_{j=1}^K \exp(\alpha_j)$.

1. Compute the partial derivatives $\frac{\partial \ln p(x)}{\partial \mu_i}$, and similar for σ_i and α_i.

Let $g(x)$ be the $3K$ dimensional gradient vector that concatenates these partial derivatives. Denote the Fisher information matrix $F = \int_x p(x) g(x) g(x)^\top$. Assume that the posteriors $p(i|x) = w_i \mathcal{N}(x; \mu_i, \sigma_i^2)/p(x)$ are sharply peaked, i.e. close to one for a single i and close to zero for all others. Decompose F into 3×3 blocks, corresponding to the w_i, μ_i and σ_i.

2. Show that F is block diagonal.

3. Show that the μ and σ blocks are diagonal, and give the diagonal entries.

Fix $\alpha_K = 0$ to remove a redundant degree of freedom from the α_i, and let $\tilde{\alpha} = (\alpha_1, \ldots, \alpha_{K-1})$. Let $\tilde{g}(x) = \nabla_{\tilde{\alpha}} \ln p(x)$ be the gradient with respect to $\tilde{\alpha}$, and similarly let $\tilde{F}$ be the Fisher information matrix with respect to $\tilde{\alpha}$.

4. Show that the Fisher kernel with respect to $\tilde{\alpha}$ can be written as $\tilde{g}(x)^\top \tilde{F}^{-1} \tilde{g}(y) = \phi(x)^\top \phi(y)$ where $\phi(x)$ is a K dimensional vector.

Exercise 3: Variational bound on marginal likelihood

Suppose the following mixture distribution $p(x) = \sum_{i=1}^{K} p(z = i)p(x|z = i)$. The entropy of a discrete distribution q is defined as $H(q) = \sum_{i=1}^{K} q_i \ln q_i$, where we use the shorthand $q_i = q(z = i)$. The Kullback Leibler divergence between distributions p and q is defined as $D(q||p) = \sum_{i=1}^{K} q_i (\ln q_i - \ln p_i)$. Assume all q_i and p_i are strictly positive.

1. Show that $F \equiv \ln p(x) - D\left(q(z)||p(z|x)\right) \leq \ln p(x)$.

2. Show that $F = H(q(z)) + \sum_{i=1}^{K} q(z = i) \left[\ln p(z = i) + \ln p(x|z = i)\right]$.

3. Show that $F = \sum_{i=1}^{K} q(z = i) \left[\ln p(x|z = i)\right] - D\left(q(z)||p(z)\right)$.

Exercise 4: Positive definite kernels

Which of these kernels are positive definite? You need to provide a proof for all cases
1. $K(x, y) = 1/(1 - xy)$ with $\mathcal{X} = (-1, 1)$;
2. $K(x, y) = \max(x, y)$ with $\mathcal{X} = [0, 1]$;
3. $K(x, y) = \cos(x + y)$ with $\mathcal{X} = \mathbb{R}$;
4. $K(x, y) = \cos(x - y)$ with $\mathcal{X} = \mathbb{R}$;
5. $K(x, y) = GCD(x, y)$ (greatest common divisor) with $\mathcal{X} = \mathbb{N}$;

Exercise 5: Kernel LDA

Fisher's linear discriminant analysis (LDA) is a method for supervised binary classification of finite-dimensional vectors. Given two sets of points $\mathcal{S}_1 = \left\{x_1^1, \ldots, x_{n_1}^1\right\}$ and $\mathcal{S}_2 = \left\{x_1^2, \ldots, x_{n_2}^2\right\}$ in $\mathbb{R}^p$, let us denote by $m_i = \frac{1}{n_i} \sum_{j=1}^{l_i} x_j^i$, and by:

$$S_B = (m_1 - m_2)(m_1 - m_2)^\top , \tag{1}$$

$$S_W = \sum_{i=1,2} \sum_{x \in \mathcal{S}_i} (x - m_i)(x - m_i)^\top , \tag{2}$$

the *between* and *within* class scatter matrices, respectively. LDA constructs the function

$$f_w(x) = w^\top x ,$$

where w is the vector which maximizes

$$J(w) = \frac{w^\top S_B w}{w^\top S_W w} .$$

1. Why does it make sense to maximize $J(w)$? What do we expect to find? (you can take as example the case where the two sets $\mathcal{S}_1$ and $\mathcal{S}_2$ form two clusters, e.g., two Gaussians).

2. We want to extend LDA to the feature space $\mathcal{H}$ induced by a positive definite kernel K by the relations $K(x, x') =< \Phi(x), \Phi(x') >_{\mathcal{H}}$. For a vector $w \in \mathcal{H}$ that is a linear combination of the form

$$w = \sum_{i=1,2} \sum_{j=1}^{n_i} \alpha_j^i \Phi(x_j^i) ,$$

express $J(w)$ and $f_w(x)$ as a function of α and K.

Exercise 6: Rademacher complexity

A Rademacher variable is a random variables σ that can take two possible values, -1 and $+1$, with equal probability $1/2$.

1. Let $(u_1, u_2, \ldots, u_N)$ be N vectors in a Hilbert space endowed with an inner product $< .,. >$, and let $\sigma_1, \sigma_2, \ldots, \sigma_N$ be N independent Rademacher variables. Show that:

$$\mathbb{E}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma_i\sigma_j < u_i, u_j >\right) = \sum_{i=1}^{N} \| u_i \|^2 .$$

2. Let K be a positive definite kernel on a space $\mathcal{X}$, $\mathcal{H}_K$ denote the associated reproducing kernel Hilbert space, and $B_R = \{f \in \mathcal{H}_K, \| f \|_{\mathcal{H}_K} \leq R\}$. Let a set of points $\mathcal{S} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ with $\mathbf{x}_i \in \mathcal{X}$ $(i = 1, \ldots, N)$, and let $\sigma_1, \sigma_2, \ldots, \sigma_N$ be N independent Rademacher variables. Show that:

$$\mathbb{E} \sup_{f \in B_R} \left| \sum_{i=1}^{N} \sigma_i f(\mathbf{x}_i) \right| \leq R \sqrt{\sum_{i=1}^{N} K(\mathbf{x}_i, \mathbf{x}_i)} .$$

Exercise 7: Conditionally positive definite kernels

Let $\mathcal{X}$ be a set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called *conditionally positive definite* (c.p.d.) if and only if it is symmetric and satisfies:

$$\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

for any $n \in \mathbb{N}$, $x_1, x_2, \ldots, x_n \in \mathcal{X}^n$ and $a_1, a_2, \ldots, a_n \in \mathbb{R}^n$ with $\sum_{i=1}^{n} a_i = 0$.

1. Show that a positive definite (p.d.) function is c.p.d.

2. Is a constant function p.d.? Is it c.p.d.?

3. If $\mathcal{X}$ is a Hilbert space, then is $k(x, y) = -||x - y||^2$ p.d.? Is it c.p.d.?

4. Let $\mathcal{X}$ be a nonempty set, and $x_0 \in \mathcal{X}$ a point. For any function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, let $\tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the function defined by:

$$\tilde{k}(x, y) = k(x, y) - k(x_0, x) - k(x_0, y) + k(x_0, x_0).$$

Show that k is c.p.d. if and only if $\tilde{k}$ is p.d.

5. Let k be a c.p.d. kernel on $\mathcal{X}$ such that $k(x, x) = 0$ for any $x \in \mathcal{X}$. Show that there exists a Hilbert space $\mathcal{H}$ and a mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such that, for any $x, y \in \mathcal{X}$,

$$k(x, y) = -||\Phi(x) - \Phi(y)||^2.$$

6. Show that if k is c.p.d., then the function $\exp(tk(x, y))$ is p.d. for all $t \geq 0$

7. Conversely, show that if the function $\exp(tk(x, y))$ is p.d. for any $t \geq 0$, then k is c.p.d.