CS 229, Spring 2023
Section #4 Solutions: MLE of Gaussian Covariance, More Kernels

1. **MLE of Gaussian Covariance Matrices**

   In this week's homework, you need to solve for the maximum likelihood estimates (MLE) of the parameters for Gaussian discrminant analysis (GDA). This involves computing the MLE of the covariance matrix Σ for a composite of functions, one of which is the multi-variate Gaussian distribution. Let's consider how we might take the gradient of just a single multi-variate Gaussian

$$p(x) = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

   with respect to its covariance Σ.

   (a) As a warm up, derive an expression (in vectorized form) for ∇_X where $f(X) = \nabla_X a^T X b$ for arbitrary vectors a, b and matrix X.

   Answer: Recall that the gradient of a function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is defined as

$$\nabla_A f(A) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

   i.e., an $m \times n$ matrix with

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}.$$

   To find $\nabla_X a^T X b$, we first find an expression for $\frac{\partial}{\partial X_{ij}} a^T X b$

$$\begin{aligned} \frac{\partial}{\partial X_{ij}} a^T X b &= \frac{\partial}{\partial X_{ij}} \sum_{i=1}^{n}\sum_{j=1}^{d} a_i b_j X_{ij} \\ &= a_i b_j \end{aligned}$$

   Thus $(\nabla_X a^T X b)_{ij} = a_i b_j$ so $\nabla_X a^T X b = ab^T$

   To compute the MLE of Σ, we consider the log-likelihood function

$$\ell = \sum_{i=1}^{n} \log p(x^{(i)}) = \sum_{i=1}^{n} -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu)$$

   and its gradient with respect to Σ. To make the problem easier, we will do a change of variables $S = \Sigma^{-1}$.

(b) Convince yourself why finding the MLE for S corresponds to finding the MLE for Σ. You can take for granted that there is indeed a maximum for the log-likelihood function at Σ. (Note you can also confirm this yourself via analysis of the Hessian, but that isn't expected for this class.)

Answer: Here's a formal approach:

Let $f(X)$ be a real-valued function defined on the domain D, i.e. $X \in D$. Let h be a one-to-one function mapping $D \to D^{-1}$ with a one-to-one inverse h^{-1} mapping $D^{-1} \to D$, so for each $X \in D$, there is a unique $X^{-1} \in D^{-1}$. Let

$$g(X^{-1}) = f(h^{-1}(X^{-1})) = f(X)$$

If the maximum of $f(X)$ is at $X = X_{max}$, then $f(X) \leq f(X_{max})$ for all $X \in D$. So

$$g(X^{-1}) = f(h^{-1}(X^{-1})) = f(X) \leq f(X_{max}) = g(h(X_{max})) = g(X_{max}^{-1})$$
$$g(X^{-1}) \leq g(X_{max}^{-1})$$

and so the maximum of $g(X^{-1})$ is at X_{max}^{-1}, corresponding to the maximum of $F(X)$ at X_{max}.

With the change of variables, we have that

$$\ell = \sum_{i=1}^{n} -\frac{k}{2}\log(2\pi) + \frac{1}{2}\log(|S|) - \frac{1}{2}(x^{(i)} - \mu)^T S (x^{(i)} - \mu)$$

This follows from the identity $|X^{-1}| = \frac{1}{|X|}$ for invertible X.

(c) Compute $\nabla_S \ell$ to find a closed form solution for the MLE of S. Then, invert this estimate to find the MLE of Σ.

Hint: The following identities (and the identity from (a)) will prove useful

$$\nabla_X |X| = |X|(X^{-1})^T$$
$$(X^{-1})^T = (X^T)^{-1}$$

Answer:

$$\nabla_S \left(\sum_{i=1}^{n} -\frac{k}{2}\log(2\pi) + \frac{1}{2}\log(|S|) - \frac{1}{2}(x^{(i)} - \mu)^T S (x^{(i)} - \mu) \right) = 0$$
$$\frac{1}{2}\sum_{i=1}^{n} \frac{1}{|S|}|S|(S^{-1})^T - (x^{(i)} - \mu)(x^{(i)} - \mu)^T = 0$$
$$\frac{1}{2}\sum_{i=1}^{n} (S^{-1} - (x^{(i)} - \mu)(x^{(i)} - \mu)^T) = 0$$

Simplifying this expression yields

$$S = \left(\frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu)(x^{(i)} - \mu)^T \right)^{-1}$$

and thus, since $S = \Sigma^{-1}$,

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

2. **Valid Kernel Functions**

Let's analyze the kernel functions from this week's homework to see if they are valid kernels. Recall from lecture the two ways to test whether a function K is a valid kernel:

- If $K(x, y) = \langle \phi(x), \phi(y) \rangle$ for a feature map $\phi(x) : \mathbb{R}^d \to \mathbb{R}^p$, then K is a valid kernel.
- (Mercer's Theorem) If the kernel matrix K is symmetric positive semi-definite, then that is necessary and sufficient for K to be a valid kernel.

Find whether the following are valid kernels:

(a) The dot product kernel: $K(x, y) = x^T y$

Answer: Yes, this function is in the form of $K(x, y) = \langle \phi(x), \phi(y) \rangle$ where $\phi(x) = x$.

(b) The indicator function $K(x, y) = \begin{cases} -1 & x = y \\ 0 & x \neq y \end{cases}$

Answer: No. Let's consider two example inputs x_1 and x_2. With just 2 examples, we can construct the 2×2 kernel matrix that contains every permutation of x_1, x_2 as arguments to K:

$$K = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) \\ K(x_2, x_1) & K(x_2, x_2) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

The eigenvalues λ can be computed from the eigenvalue problem $\det(K - \lambda I) = (-1 - \lambda)^2 - 0^2 = 0 \to \lambda = -1$. All the eigenvalues of a PSD matrix have to be nonnegative, so K is not PSD and hence not a valid kernel by Mercer's theorem.

(c) The radial basis function (RBF) kernel: $K(x, y) = \exp\left(-\frac{||x-y||_2^2}{2\sigma^2}\right)$

Answer: Some key facts first:

- The sum of 2 (symmetric) PSD matrices is (symmetric) PSD.
- The elementwise (aka Hadamard) product of 2 (symmetric) PSD matrices is (symmetric) PSD, i.e. if K_1 and K_2 are PSD, then the matrix K_3 consisting of entries $K_3(x, y) = K_1(x, y)K_2(x, y)$ is PSD.

Rewrite

$$\begin{aligned} K(x, y) &= \exp\left(-\frac{||x - y||_2^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{x^T x}{2\sigma^2}\right)\exp\left(\frac{x^T y}{\sigma^2}\right)\exp\left(-\frac{y^T y}{2\sigma^2}\right) \end{aligned}$$

Notice how the expression within the middle exponent is a valid kernel $K'(x, y) \propto x^T y$. Consider the Taylor's expansion

$$\exp\left(K'(x,y)\right) = \sum_{n=0}^{\infty} \frac{1}{n!} K'(x,y)^n$$

which is just a summation of elementwise products of valid kernel K', hence exponentiation of a valid kernel also produces a valid kernel.

Rewrite

$$\begin{aligned} K(x,y) &= \exp\left(-\frac{x^T x}{2\sigma^2}\right) \exp\left(\frac{x^T y}{\sigma^2}\right) \exp\left(-\frac{y^T y}{2\sigma^2}\right) \\ &= f(x) K^*(x,y) f(y) \end{aligned}$$

letting K^* be the kernel that represents our middle exponentiation expression and $f : \mathbb{R}^d \to \mathbb{R}$ represent our left and right exponentiation expressions. Since K^* is a valid kernel, then it can be expressed as $K^*(x,y) = \langle \phi^*(x), \phi^*(y) \rangle$, thus

$$\begin{aligned} K(x,y) &= f(x) K^*(x,y) f(y) \\ &= f(x) \langle \phi^*(x), \phi^*(y) \rangle f(y) \\ &= \langle \phi'(x), \phi'(y) \rangle \end{aligned}$$

where $\phi'(x) = \phi^*(x) f(x)$, noting that $f(x)^T = f(x)$ since it is a scalar. Hence, K is a valid kernel.