CS 229, Spring 2023
Section #5 Solutions: Review Kernels, GLM

---

1. **Valid Kernels**

   Continuing our discussion of valid kernels from Discussion Section #4, state whether the following are valid kernel functions and why. Recall that:

   - 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.

   (a) $K(x, y) = (1 + x^T y)^2 + x^T y$

   Answer: Yes. This is a summation of a polynomial kernel and a dot product kernel, each of which are valid kernel functions. The summation of two valid kernels leads to a valid kernel due to Mercer's Theorem, since the summation of two symmetric PSD matrices is also symmetric PSD.

   (b) $K(x, y) = (1 + x^T y)^2 - x^T y$

   Answer: Not necessarily. The difference between two symmetric PSD matrices is not always symmetric PSD. To see this, consider how for an arbitrary vector z, we have $z^T K_1 z \geq 0$ and $z^T K_2 z \geq 0$ for two valid kernels K_1, K_2, but $z^T K_1 z - z^T K_2 z$ is not guaranteed to be greater than or equal to 0.

   (c) $K(x, y) = \exp((1 + x^T y)^2 + x^T y)$

   Answer: Yes. As shown in Section #4 solutions, exponentiation of a valid kernel leads to a valid kernel.

   (d) $K(x, y) = \dfrac{(1 + x^T y)^2 + x^T y}{c}$ for some constant $c \in \mathbb{R}, c \neq 0$

   Answer: Not necessarily. Let $K'(x, y)$ equal to the numerator be a valid kernel. If $c < 0$, then for arbitrary vector z, we have $z^T \frac{K'(x,y)}{c} z \leq 0$, which is not PSD.

   (e) $K(x, y) = \exp\left(\dfrac{x^T x}{c}\right) \exp\left(\dfrac{(1 + x^T y)^2 + x^T y}{c^2}\right) \exp\left(\dfrac{y^T y}{c}\right)$ for some constant $c \in \mathbb{R}, c \neq 0$

   Answer: Yes. As shown in Section #4 solutions, a function in the form of $K(x, y) = f(x)K'(x, y)f(y)$ where K' is a valid kernel leads to a valid kernel.

2. **Log-Normal GLM**

   In Problem Set 1, you worked with a generalized linear model (GLM) that utilized the Poisson distribution. In that problem, you:

   - showed that the Poisson distribution is in the exponential family, i.e. it has the form

   $$p(y; \eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

   - derived the stochastic gradient descent update rule for the model by taking the gradient of the log-likelihood function $\log p(y^{(i)}|x^{(i)}; \theta)$ (of an example) with respect to θ.

Here, we'll examine another GLM that uses the log-normal distribution parameterized by $\mu \in \mathbb{R}$ for $y \in \mathbb{R}$:

$$p(y;\mu) = \frac{1}{y\sqrt{2\pi}}\exp\left(-\frac{1}{2}(\log(y)-\mu)^2\right)$$

(a) Show that the log-normal distribution is in the exponential family. What are $b(y), \eta, T(y), a(\eta)$?

Answer: Rewrite the distribution as:

$$p(y;\mu) = \frac{1}{y\sqrt{2\pi}}\text{exp}\left(-\frac{1}{2}(\text{log}^2(y))\right)\text{exp}\left(\mu\text{log}(y) - \frac{1}{2}\mu^2\right)$$

We can then let:

$$\begin{aligned}
b(y) &= \frac{1}{y\sqrt{2\pi}}\text{exp}\left(-\frac{1}{2}(\text{log}^2(y))\right) \\
\eta &= \mu \\
T(y) &= \text{log}(y) \\
a(\eta) &= \frac{1}{2}\mu^2, \text{ since we let } \eta = \mu
\end{aligned}$$

(b) Derive the stochastic gradient descent update rule with learning rate α for a GLM model that utilizes the above log-normal distribution for its data: $\{(x^i, y^i)\}; i = 1, ..., n$. Recall that for a GLM, we assume $\eta = \theta^T x$.

Answer: First, write down the log-likelihood function:

$$\begin{aligned}
l(\theta) &= \sum_{i=1}^{n} \text{log } p(y^{(i)}|x^{(i)};\theta) \\
&= \sum_{i=1}^{n} \text{log } \left(\frac{1}{y^{(i)}\sqrt{2\pi}}\text{exp}\left(-\frac{1}{2}(\text{log}(y^{(i)}) - \theta^T x^{(i)})^2\right)\right) \\
&= \sum_{i=1}^{n} \text{log } \frac{1}{y^{(i)}\sqrt{2\pi}} - \frac{1}{2}\sum_{i=1}^{n}(\text{log}(y^{(i)}) - \theta^T x^{(i)})^2
\end{aligned}$$

Take the gradient of the log-likelihood with respect to θ:

$$\begin{aligned}
\frac{\partial \text{ log } p(y^{(i)}|x^{(i)};\theta)}{\partial\theta} &= \frac{\partial\left(\text{ log } \frac{1}{y^{(i)}\sqrt{2\pi}} - \frac{1}{2}(\text{log}(y^{(i)}) - \theta^T x^{(i)})^2\right)}{\partial\theta} \\
&= \frac{\partial\left(-\frac{1}{2}(\text{log}(y^{(i)}) - \theta^T x^{(i)})^2\right)}{\partial\theta} \\
&= -(\text{log}(y^{(i)}) - \theta^T x^{(i)})x^{(i)}
\end{aligned}$$

And so our update rule is:

$$\theta := \theta + \alpha(\text{log}(y^{(i)}) - \theta^T x^{(i)})x^{(i)}$$