Labs
Machine Learning Course
Fall 2025

EPFL
School of Computer and Communication Sciences
Bob West
epfml.github.io/cs433-2025/

Problem Set 7, Oct 30, 2025 (Solutions to Theory Questions Part)

Problem 1 (Kernels):

In class we have seen that many kernel functions $k(\boldsymbol{x}, \boldsymbol{x}')$ can be written as inner products $\phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$, for a suitably chosen vector-function $\phi(\cdot)$ (often called a feature map). Let us say that such a kernel function is *valid*. We further discussed many operations on valid kernel functions that result again in valid kernel functions. Here are two more.

1. Let $k_1(\boldsymbol{x}, \boldsymbol{x}')$ be a valid kernel function. Let f be a polynomial with positive coefficients. Show that $k(\boldsymbol{x}, \boldsymbol{x}') = f(k_1(\boldsymbol{x}, \boldsymbol{x}'))$ is a valid kernel.
2. Show that $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(k_1(\boldsymbol{x}, \boldsymbol{x}'))$ is a valid kernel assuming that $k_1(\boldsymbol{x}, \boldsymbol{x}')$ is a valid kernel. *Hint*: You can use the following property: if $(K_n)_{n\geq 0}$ is a sequence of valid kernels and if there exists a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for all $(x, x') \in \mathcal{X}^2$, $K_n(x, x') \underset{n\to+\infty}{\longrightarrow} K(x, x')$, then K is a valid kernel.

Solution: *Main Idea.* For (1), we use closure of valid kernels under sums, products, and scaling by $c \geq 0$; any polynomial with nonnegative coefficients is built from these operations applied to k_1, hence remains valid. For (2), we use the Taylor series $\exp(k_1) = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{1}{i!} k_1^i$ together with closure under pointwise limits of valid kernels.

1.
   - First we will prove that the sum or two valid kernels k_1 and k_2 $k = k_1 + k_2$ is a valid kernel. We need to construct a feature vector $\phi(\boldsymbol{x})$ such that $k(\boldsymbol{x}, \boldsymbol{x}') = \phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}')$, then by definition k would be a valid kernel.
   Because kernels k_1 and k_2 are valid kernels
   $$k_1(\boldsymbol{x}, \boldsymbol{x}') = \phi_1(\boldsymbol{x})^\top \phi_1(\boldsymbol{x}'), \qquad k_2(\boldsymbol{x}, \boldsymbol{x}') = \phi_2(\boldsymbol{x})^\top \phi_2(\boldsymbol{x}'),$$
   for some feature vectors $\phi_1(\boldsymbol{x})$ and $\phi_2(\boldsymbol{x})$.
   Lets take $\phi(\boldsymbol{x}) = \begin{pmatrix} \phi_1(\boldsymbol{x}) \\ \phi_2(\boldsymbol{x}) \end{pmatrix}$, then
   $$\begin{aligned} \phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}') = \left(\phi_1(\boldsymbol{x})^\top, \phi_2(\boldsymbol{x})^\top\right) \begin{pmatrix} \phi_1(\boldsymbol{x}') \\ \phi_2(\boldsymbol{x}') \end{pmatrix} &= \phi_1(\boldsymbol{x})^\top \phi_1(\boldsymbol{x}') + \phi_2(\boldsymbol{x})^\top \phi_2(\boldsymbol{x}') \\ &= k_1(\boldsymbol{x}, \boldsymbol{x}') + k_2(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}') \end{aligned}$$
   Therefore $k = k_1 + k_2$ is a valid kernel.
   - Second, we will prove that the product $k = k_1 \cdot k_2$ of two valid kernels is a valid kernel.
   Let's denote n_1 and n_2 dimensions of a feature vectors $\phi_1(\boldsymbol{x})$ and $\phi_2(\boldsymbol{x})$ (i.e. $\phi_1(\boldsymbol{x}) \in \mathbb{R}^{n_1}$, $\phi_2(\boldsymbol{x}) \in \mathbb{R}^{n_1}$).
   $$k_1(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=0}^{n_1-1} \phi_{1,i}(\boldsymbol{x})\phi_{1,i}(\boldsymbol{x}'), \qquad k_2(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=0}^{n_2-1} \phi_{2,j}(\boldsymbol{x})\phi_{2,j}(\boldsymbol{x}'),$$
   Then the kernel $k = k_1 \cdot k_2$ is
   $$k(\boldsymbol{x}, \boldsymbol{x}') = \left(\sum_{i=0}^{n_1-1} \phi_{1,i}(\boldsymbol{x})\phi_{1,i}(\boldsymbol{x}')\right) \left(\sum_{j=0}^{n_2-1} \phi_{2,j}(\boldsymbol{x})\phi_{2,j}(\boldsymbol{x}')\right) = \sum_{i=0}^{n_1-1} \sum_{j=0}^{n_2-1} \left(\phi_{1,i}(\boldsymbol{x})\phi_{2,j}(\boldsymbol{x})\right)\left(\phi_{1,i}(\boldsymbol{x}')\phi_{2,j}(\boldsymbol{x}')\right)$$

Lets introduce a feature vector $\phi(\boldsymbol{x}) \in \mathbb{R}^{n_1 n_2}$, such that $\phi_{in_2+j}(\boldsymbol{x}) = \phi_{1,i}(\boldsymbol{x})\phi_{2,j}(\boldsymbol{x})$ for $i \in [0, \ldots, n_1 - 1], j \in [0, \ldots, n_2 - 1]$. Note that for such i and j the index of the feature vector ϕ ic correct: $in_2 + j \in [0, \ldots, n_1 n_2 - 1]$. Then,

$$
\begin{aligned}
\phi(\boldsymbol{x})^\top \phi(\boldsymbol{x}') &= \sum_{l=0}^{n_1 n_2 - 1} \phi_l(\boldsymbol{x})\phi_l(\boldsymbol{x}') = \sum_{i=0}^{n_1-1}\sum_{j=0}^{n_2-1} \phi_{in_2+j}(\boldsymbol{x})\phi_{in_2+j}(\boldsymbol{x}') \\
&= \sum_{i=0}^{n_1-1}\sum_{j=0}^{n_2-1} \left(\phi_{1,i}(\boldsymbol{x})\phi_{2,j}(\boldsymbol{x})\right)\left(\phi_{1,i}(\boldsymbol{x}')\phi_{2,j}(\boldsymbol{x}')\right) = k_1(\boldsymbol{x}, \boldsymbol{x}') \cdot k_2(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}, \boldsymbol{x}').
\end{aligned}
$$

Therefore $k(\boldsymbol{x}, \boldsymbol{x}') = k_1(\boldsymbol{x}, \boldsymbol{x}') \cdot k_2(\boldsymbol{x}, \boldsymbol{x}')$ is a valid kernel.

- Third we need to show that if k_1 is a valid kernel, then $k = c \times k_1$ with $c \geq 0$ is also a valid kernel. Since k_1 is valid, we can write for all x and x': $k_1(x, x') = \phi_1(x)^T\phi_1(x')$. Now let $\phi(x) = \sqrt{c}\phi_1(x)$. Notice that $k(x, x') = c \cdot k_1(x, x') = (\sqrt{c}\phi_1(x))^T(\sqrt{c}\phi_1(x')) = \phi(x)^T\phi(x')$. Hence $c \times k_1$ is also a valid kernel.
- Since f is only composed of the three previous operations (sum, product, multiplication by positive scalar), we can conclude that $(x, x') \mapsto f(k_1(x, x'))$ is a valid kernel.

2. It suffices to apply the hint to the sequence of kernels $k_n(x, x') = \sum_{i=0}^{n} \frac{1}{i!} k_1(x, x')^i$. According to our previous result these are valid kernels. Since $k_n(x, x') \underset{n \to +\infty}{\longrightarrow} \exp(k_1(x, x')$ we can apply the hint and conclude that k is a valid kernel.

Bonus. For the curious who are familiar with matrices and the trace operator, here is an elegant and more natural way of showing that the product of two valid kernels is a valid kernel. Notice that for $x, x' \in \mathcal{X}^2$:

$$
\begin{aligned}
k(x, x') &= k_1(x, x') \cdot k_2(x, x') \\
&= \phi_1(x)^T\phi_1(x')\phi_2(x)^T\phi_2(x') \\
&= \phi_1(x)^T\phi_1(x')\phi_2(x')^T\phi_2(x) \\
&= \text{trace}(\phi_1(x)^T\phi_1(x')\phi_2(x')^T\phi_2(x)) \\
&= \text{trace}(\phi_1(x')\phi_2(x')^T\phi_2(x)\phi_1(x)^T) \\
&= \text{trace}((\phi_2(x')\phi_1(x')^T)^T\phi_2(x)\phi_1(x)^T) \\
&= \langle \phi_2(x)\phi_1(x)^T,\ \phi_2(x')\phi_1(x')^T \rangle_F
\end{aligned}
$$

Second equality is the definition of the valid kernels k_1 and k_2, third is due to $x^Ty = y^Tx$, fourth is noticing that for $z \in \mathbb{R}$ trace$(z) = z$, fifth is that trace$(AB) =$ trace(BA), seventh is due to $xy^T = (yx^T)^T$, and last is the definition of the Frobenius inner product for matrices. Hence by letting $\phi(x) = \phi_2(x)\phi_1(x)^T \in \mathbb{R}^{n_2 \times n_1}$ we obtain $k(x, x') = \langle \phi(x), \phi(x') \rangle_F$.

Problem 2 (Softmax Cross Entropy):

In the notebook exercises we performed multiclass classification using softmax-cross-entropy as our loss. The softmax of a vector $\mathbf{x} = [x_1, ..., x_d]^\top$ is a vector $\mathbf{z} = [z_1, ..., z_d]^\top$ with:

$$z_k = \frac{\exp(x_k)}{\sum_{i=1}^{d} \exp(x_i)} \tag{1}$$

The label y is an integer denoting the target class. To turn y into a probability distribution for use with cross-entropy, we use one-hot encoding:

$$\text{onehot}(y) = \mathbf{y} = [y_1, ..., y_d]^\top \text{ where } y_k = \begin{cases} 1, & \text{if } k = y \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The cross-entropy is given by:

$$H(\mathbf{y}, \mathbf{z}) = -\sum_{i=1}^{d} y_i \ln(z_i) \tag{3}$$

We ask you to do the following:

1. Equation 1 potentially computes $\exp$ of large positive numbers which is numerically unstable. Modify Eq. 1 to avoid positive numbers in $\exp$. Hint: Use $\max_j(x_j)$.
2. Derive $\frac{\partial H(\mathbf{y},\mathbf{z})}{\partial x_j}$. You may assume that $\mathbf{y}$ is a one-hot vector.
3. What values of x_i minimize the softmax-cross-entropy loss? To avoid complications, practitioners sometimes use a trick called label smoothing where $\mathbf{y}$ is replaced by $\hat{\mathbf{y}} = (1-\epsilon)\mathbf{y} + \frac{\epsilon}{d}\mathbf{1}$ for some small value e.g. $\epsilon = 0.1$.

Solution:

Part 1:

$$z_k = \frac{\exp(x_k)}{\sum_{i=1}^d \exp(x_i)} = \frac{\exp(-\max_j(x_j))}{\exp(-\max_j(x_j))}\frac{\exp(x_k)}{\sum_{i=1}^d \exp(x_i)} = \frac{\exp(x_k - \max_j(x_j))}{\sum_{i=1}^d \exp(x_i - \max_j(x_j))} \tag{4}$$

Part 2:

$$\begin{aligned}\frac{\partial H(\mathbf{y},\mathbf{z})}{\partial x_j} &= \frac{\partial H(\mathbf{y},\mathbf{z})}{\partial z_y}\frac{\partial z_y}{\partial x_j} \\ &= \frac{-1}{z_y}\frac{\partial}{\partial x_j}\frac{\exp(x_y)}{\sum_{i=1}^d \exp(x_i)}\end{aligned}$$

For $j = y$ we have:

$$\frac{-1}{z_y}\frac{\partial}{\partial x_j}\frac{\exp(x_j)}{\sum_{i=1}^d \exp(x_i)} = -\frac{\sum_{i=1}^d \exp(x_i)}{\exp(x_j)}\cdot\frac{\exp(x_j)\sum_{i=1}^d \exp(x_i) - \exp(x_j)^2}{(\sum_{i=1}^d \exp(x_i))^2} = -\frac{\sum_{i=1}^d \exp(x_i) - \exp(x_j)}{\sum_{i=1}^d \exp(x_i)} = z_j - 1$$

For $j \neq y$ we have:

$$\frac{-1}{z_y}\frac{\partial}{\partial x_j}\frac{\exp(x_y)}{\sum_{i=1}^d \exp(x_i)} = -\frac{\sum_{i=1}^d \exp(x_i)}{\exp(x_y)}\cdot\frac{-\exp(x_j)\exp(x_y)}{(\sum_{i=1}^d \exp(x_i))^2} = \frac{\exp(x_j)}{\sum_{i=1}^d \exp(x_i)} = z_j$$

We can concisely write:

$$\frac{\partial H(\mathbf{y},\mathbf{z})}{\partial \mathbf{x}} = \mathbf{z} - \mathbf{y} \tag{5}$$

Part 3: The optimality condition based on setting the gradient to 0 suggests that $\mathbf{z} - \mathbf{y} = 0$. This means that when j is equal to the correct label y, we must have $z_j = y_j$. Since $z_j = softmax(\mathbf{x})_j$ and $y_j = 1$ in this case, this implies that the following must hold:

$$\frac{e^{x_j}}{\sum_{i=1}^d e^{x_i}} = 1.$$

This can be rewritten as:

$$\frac{\sum_{i=1}^d e^{x_i}}{e^{x_j}} = 1 \implies \sum_{i=1}^d e^{x_i - x_j} = 1 \implies e^{x_j - x_j} + \sum_{i\neq j} e^{x_i - x_j} = 1 \implies \sum_{i \neq j} e^{x_i - x_j} = 0.$$

From this, we conclude that for all i not equal to j, we must have $e^{x_i - x_j} \to 0$ or, equivalently, $x_i - x_j \to -\infty$. This means that $x_i \to -\infty$ (again, for all i not equal to j, i.e., for $i \neq y$) and $x_j \to \infty$ as suggested in the solution. Also, one can verify that this solution is also consistent with the optimality of x_j for $j \neq y$. Thus, we conclude that the loss is minimized when $x_j \to \begin{cases}\infty \text{ for } j = y \\ -\infty \text{ else}\end{cases}$.

Note that the expression $\frac{\partial \mathbf{H}(\mathbf{y},\mathbf{z})}{\partial \mathbf{x}} = \mathbf{z} - \mathbf{y}$ is true only if $\mathbf{y}$ is a one-hot vector. With label smoothing, $\mathbf{y}$ becomes a "smoothed" version of the one-hot vector, and thus we would first need to derive a different expression for the derivative of the cross-entropy loss. After doing that, it should become apparent that the optimal softmax values $\mathbf{z}$ shouldn't be a one-hot vector anymore which will correspond to a finite minimum in terms of $\mathbf{x}$. Thus, the main effect of label smoothing is that it makes the minimum finite.