CS5487 Problem Set 10
Kernels

Antoni Chan
Department of Computer Science
City University of Hong Kong

Kernel functions

Problem 10.1 Constructing kernels from kernels

Suppose $k_1(x,z)$ and $k_2(x,z)$ are valid kernels. Prove the following kernels are also valid kernels:

(a) kernel scaling:	$k(x,z) = ck_1(x,z)$,	where $c > 0$.
(b) sum:	$k(x,z) = k_1(x,z) + k_2(x,z)$.	
(c) product:	$k(x,z) = k_1(x,z)k_2(x,z)$.	
(d) input scaling:	$k(x,z) = f(x)k_1(x,z)f(z)$,	for any function $f(\cdot)$.
(e) polynomial:	$k(x,z) = k_1(x,z)^q$,	where q is a positive integer.
(f) exponential:	$k(x,z) = \exp(k_1(x,z))$.	

. . . . . .

Problem 10.2 Kernels on real vectors

Let $x, z \in \mathbb{R}^n$. Show the following are valid kernels:

(a) Gaussian or RBF:	$k(x,z) = \exp(-\alpha \|x-z\|^2)$,	for $\alpha > 0$.
(b) exponential:	$k(x,z) = \exp(-\alpha \|x-z\|)$,	for $\alpha > 0$.
(c) linear:	$k(x,z) = x^T A z$,	where A is a positive definite matrix.
(d) polynomial:	$k(x,z) = (x^T z + 1)^q$,	where q is a positive integer.
(e) cosine:	$k(x,z) = \frac{x^T z}{\|x\|\|z\|}$.	
(f) periodic:	$k(x,z) = \exp\{-\alpha \sin^2(\frac{x-z}{2})\}$	$x, z \in \mathbb{R}$.

For (f), consider the warping $u(x) = [\sin(x), \cos(x)]^T$ and the fact that $\|u(x) - u(z)\|^2 = 4\sin^2(\frac{x-z}{2})$. What are the feature transformations corresponding to each of these kernels?

. . . . . .

Problem 10.3 Constructing kernels from kernels (part 2 – feature selection)

Suppose $x, z \in \mathbb{R}^d$, $\phi(x)$ is a function from x to $\mathbb{R}^M$, and $k_3(\cdot,\cdot)$ is a valid kernel in $\mathbb{R}^M$. Prove the following is a valid kernel:

(a) $k(x,z) = k_3(\phi(x), \phi(z))$.

Let x_a and x_b be variables with $x = (x_a, x_b)$, and k_a and k_b are valid kernels over their respective spaces. Show that the following are valid kernels:

(b) $k(x,z) = k_a(x_a, z_a) + k_b(x_b, z_b)$.
(c) $k(x,z) = k_a(x_a, z_a)k_b(x_b, z_b)$.

. . . . . .

Problem 10.4 Constructing kernels from kernels (part 3 – normalization)

Some kernels are poorly scaled, in the sense that the dynamic range of the values is much less than the absolute values. One solution to this problem is to "normalize" the kernel,

$$\tilde{k}(x,z) = \frac{k(x,z)}{\sqrt{k(x,x)k(z,z)}}. \tag{10.1}$$

This will set the self-similarity value to $k(x,x) = 1$.

(a) Show that the normalized kernel $\tilde{k}$ is a valid kernel.

(b) Let $\Phi(x)$ be the feature transformation associated with the kernel k. Show that $\tilde{k}$ is equivalent to calculating the cosine between $\Phi(x)$ and $\Phi(z)$ in the high-dimensional feature space.

(c) What is the range of possible values of the normalized kernel $\tilde{k}(x,z)$?

.....

Problem 10.5 Constructing kernels from kernels (part 4 – kernel distance)

If $k(x,z)$ is a valid kernel, then it can be interpreted as a dot product. Hence, we can also construct a distance (norm) based on this dot product.

$$\|x-z\|^2 = x^Tx - 2x^Tz + z^Tz \quad \Rightarrow \quad d^2(x,z) = k(x,x) - 2k(x,z) + k(z,z). \tag{10.2}$$

This squared-distance can be used in place the standard L2 norm.

(a) Show that the exponential kernel distance is a valid kernel:

$$k(x,z) = \exp\{-d(k(x,x) - 2k(x,z) + k(z,z))\}. \tag{10.3}$$

This is similar to a Gaussian kernel, but with the L2-norm replaced with the kernel distance.

.....

Problem 10.6 Kernels for histograms

Let x and z be d-dimensional histograms, i.e., $x, z \in \mathbb{R}^d$, where each bin $x_i, z_i \geq 0$. Show that the following kernels between histograms are valid kernels:

(a) correlation: $k(x,z) = \sum_{i=1}^d x_i z_i$.
(b) Bhattacharyya: $k(x,z) = \sum_{i=1}^d \sqrt{x_i}\sqrt{z_i}$.
(c) histogram intersection: $k(x,z) = \sum_{i=1}^d \min(x_i, z_i)$.
(d) χ^2-RBF: $k(x,z) = \exp\left\{-\alpha\chi^2(x,z)\right\}$, $\chi^2(x,z) = \sum_{i=1}^d \frac{(x_i-z_i)^2}{\frac{1}{2}(x_i+z_i)}$.

.....

Problem 10.7 Kernels on sets

Let $X = \{x_1, \cdots, x_M\}$ and $Z = \{z_1, \cdots, z_P\}$ be set, where x, z are the set items (these could be real vectors or symbols). Define a kernel between items as $k(x, z)$.

(a) Show that the kernel between sets X and Z,

$$\tilde{k}(X, Z) = \sum_{x \in X} \sum_{z \in Z} k(x, z), \tag{10.4}$$

is a valid kernel. Hint: consider the feature transformation $\Phi(X) = \sum_{i=1}^{M} \phi(x_i)$, where $\phi(x_i)$ is the feature map induced by $k(x, z)$.

(b) Show that the kernel

$$\tilde{k}(X, Z) = \frac{1}{|X||Z|} \sum_{x \in X} \sum_{z \in Z} k(x, z)^q, \tag{10.5}$$

is a valid kernel, where q is a non-negative integer. This is called an *exponent match kernel*.

(c) Show that the kernel, which counts the number of common elements, is a valid kernel,

$$\tilde{k}(X, Z) = |X \cap Z|. \tag{10.6}$$

(d) Show that the following kernel is also valid.

$$\tilde{k}(X, Z) = 2^{|X \cap Z|}. \tag{10.7}$$

Another kernel on sets is the *pyramid match kernel*. There are also other kernels that are not positive definite, which have the form

$$k(X, Z) = e^{-\alpha d(X,Z)^2} \tag{10.8}$$

using the sum-of-minimum distances

$$d(X, Z) = \frac{1}{|X| + |Z|} \left\{ \sum_i \min_j d(x_i, z_j) + \sum_i \min_j d(z_i, x_j) \right\} \tag{10.9}$$

or the max-of-minimum distances (Haussdorf distance),

$$d(X, Z) = \max(\max_i \min_j d(x_i, z_j), \max_i \min_j d(z_i, x_j)). \tag{10.10}$$

. . . . . .

Problem 10.8 Kernels on distributions

Consider that the two inputs are distributions, $p(x)$ and $q(x)$. The correlation is a valid kernel between distributions

$$k(p, q) = \int p(x) q(x) dx. \tag{10.11}$$

(a) Show that the kernel

$$k(p, q) = \int p(x)^\rho q(x)^\rho dx \tag{10.12}$$

is a valid kernel, for any $\rho > 0$. This is called the *probability product kernel*. A special case is the Bhattacharyya kernel ($\rho = 0.5$), and the correlation kernel ($\rho = 1$).

. . . . . .

Problem 10.9 Kernels on sample spaces

Let $\mathcal{X}$ be a sample space of random variable x, and A and B events, where

$$p(A) = \int_{x \in A} p(x)dx \tag{10.13}$$

define the kernel

$$k : \mathcal{X} \times \mathcal{X} \to [-1, 1] \tag{10.14}$$

$$k(A, B) = p(A \cap B) - p(A)p(B) \tag{10.15}$$

(a) Show that $k(A, B)$ is a valid kernel. Hint: consider the feature transformation $\Phi(A) = 1_A - p(A)$, where 1_A is the function that takes the value 1 on the set A, and 0 otherwise.

.....

Kernel trick

Problem 10.10 Kernel SVM bias term

The bias term for the linear SVM can be calculated as

$$b^* = \frac{1}{|SV|} \sum_{i \in SV} (y_i - w^T x_i) \tag{10.16}$$

where $SV = \{i | \alpha_i > 0\}$ is the set of support vectors.

(a) Derive an expression for b^* for the kernel SVM.

.....

Problem 10.11 Kernel logistic regression

Consider the two-class *regularized* logistic regression from Problem 8.4, which is interpretable as MAP estimation. The prior distribution on w is zero-mean Gaussian with known precision matrix Γ (i.e., inverse of the covariance matrix),

$$p(w) = \mathcal{N}(w|0, \Gamma^{-1}). \tag{10.17}$$

Given the training set $\mathcal{D} = \{X, y\}$, the MAP estimate is

$$w^* = \underset{w}{\operatorname{argmax}} \log p(y|X, w) + \log p(w), \tag{10.18}$$

which can be calculated using the Newton-Raphson method, with the iterations

$$w^{(new)} = (XRX^T + \Gamma)^{-1} XRz, \tag{10.19}$$

where R and z are calculated from the previous $w^{(old)}$,

$$\pi_i = \sigma(x_i^T w^{(\text{old})}), \quad \pi = [\pi_1, \cdots, \pi_n]^T, \tag{10.20}$$

$$R = \text{diag}(\pi_1(1 - \pi_1), \cdots, \pi_n(1 - \pi_n)), \tag{10.21}$$

$$z = X^T w^{(old)} - R^{-1}(\pi - y). \tag{10.22}$$

$\sigma(a) = \frac{1}{1+e^{-a}}$ is the logistic sigmoid function. In this problem, we will *kernelize* the MAP version of logistic regression to obtain *kernel logistic regression*.

(a) Define $\alpha_* = w^T x_*$ as the linear function of a novel input x_*. Use the form in (10.19) to show that the linear function $w^T x_*$ can be kernelized when z is known,

$$\alpha_* = w^T x_* = k_*^T (K + R^{-1})^{-1} z \tag{10.23}$$

where $k_* = [k(x_*, x_1), \cdots, k(x_*, x_n)]^T$ is the test kernel, and $K = [k(x_i, x_j)]_{ij}$ is the training kernel matrix. Hence, the probability of class 1 for a new point x_* is $\pi_* = \sigma(\alpha_*)$. Hint: use a matrix inverse identity (Problem 1.15), and note that $x_i^T \Gamma^{-1} x_j$ is an inner product.

(b) Using (10.23), show that the z in (10.22) can be calculated iteratively from $z^{(\text{old})}$,

$$\alpha^{(\text{old})} = K(K + R^{-1})^{-1} z^{(\text{old})} \tag{10.24}$$

$$z = \alpha^{(\text{old})} - R^{-1}(\pi - y), \tag{10.25}$$

where $\pi_i = \sigma(\alpha_i^{(\text{old})})$. Hence, the Newton-Raphson iterations can also be kernelized for training.

(c) What happened to the prior covariance Γ^{-1}? What is the relationship between the prior and the kernel function, e.g., when $\Gamma = \lambda I$?

(d) Is the scale of the kernel important? In other words, is using $k(x_i, x_j)$ equivalent to using $\hat{k}(x_i, x_j) = \beta k(x_i, x_j)$, for some $\beta > 0$? Why or why not?

.....

Problem 10.12 Gaussian process regression — nonlinear Bayesian regression

In this problem we will apply the "kernel trick" to Bayesian linear regression (Problem 3.10) to obtain a non-linear Bayesian regression algorithm, known as *Gaussian process regression*. The prior distribution on the linear weights is Gaussian $p(\theta) = \mathcal{N}(\theta|0, \Gamma)$. The posterior distribution of the parameters is Gaussian

$$p(\theta|\mathcal{D}) = \mathcal{N}(\theta|\hat{\mu}_\theta, \hat{\Sigma}_\theta), \tag{10.26}$$

$$\hat{\mu}_\theta = (\Gamma^{-1} + \Phi\Sigma^{-1}\Phi^T)^{-1}\Phi\Sigma^{-1} y, \tag{10.27}$$

$$\hat{\Sigma}_\theta = (\Gamma^{-1} + \Phi\Sigma^{-1}\Phi^T)^{-1}, \tag{10.28}$$

where $\hat{\mu}_\theta$ is the posterior mean and $\hat{\Sigma}_\theta$ is the posterior covariance. Given a novel input x_*, the predictive distribution of the corresponding output $f_* = f(x_*, \theta)$ is also Gaussian

$$p(f_*|x_*, \mathcal{D}) = \mathcal{N}(f_*|\hat{\mu}_*, \hat{\sigma}_*^2), \tag{10.29}$$

$$\hat{\mu}_* = \phi(x_*)^T \hat{\mu}_\theta, \tag{10.30}$$

$$\hat{\sigma}_*^2 = \phi(x_*)^T \hat{\Sigma}_\theta \phi(x_*), \tag{10.31}$$

where the posterior mean $\hat{\mu}_\theta$ and covariance $\hat{\Sigma}_\theta$ are given in (10.27) and (10.28). We will assume that the observation noise is i.i.d, i.e., $\Sigma = \sigma^2 I$.

(a) Show that the predictive mean $\hat{\mu}_*$ can be written in the form,

$$\hat{\mu}_* = \phi_*^T \Gamma\Phi(\Phi^T\Gamma\Phi + \sigma^2 I)^{-1} y, \tag{10.32}$$

where $\phi_* = \phi(x_*)$. Hint: use a matrix inverse identity from Problem 1.15.

(b) Show that the predictive variance $\hat{\sigma}_*^2$ can be written in the form,

$$\hat{\sigma}_*^2 = \phi_*^T \Gamma \phi_* - \phi_*^T \Gamma \Phi (\Phi^T \Gamma \Phi + \sigma^2 I)^{-1} \Phi \Gamma \phi_*. \tag{10.33}$$

Hint: use a matrix inverse identity from Problem 1.15.

(c) The "kernel trick" can be applied to (a) and (b), by defining setting $k(x_i, x_j) = \phi(x_i)^T \Gamma \phi(x_j)$ yielding,

$$\hat{\mu}_* = k_*^T (K + \sigma^2 I)^{-1} y \tag{10.34}$$

$$\hat{\sigma}_*^2 = k_{**} - k_*^T (K + \sigma^2 I)^{-1} k_*, \tag{10.35}$$

where $K = [k(x_i, x_j)]_{ij}$ is the kernel matrix, $k_* = [k(x_*, x_i)]_i$ is the test kernel vector, and $k_{**} = k(x_*, x_*)$. What happened to the prior covariance Γ? What is the relationship between the prior covariance Γ and the kernel function $k(x_i, x_j)$?

(d) Define $z = (K + \sigma^2 I)^{-1} y$. The predictive mean $\hat{\mu}_*$, which is a function of x_*, can then be written as

$$\hat{\mu}(x_*) = \sum_{i=1}^{n} z_i k(x_*, x_i), \tag{10.36}$$

where z_i is the ith element of z. Hence, the regressed function is a linear combination of kernel functions $k(x_*, x_i)$, where x_i is fixed. What will be the general shape of the regressed function for the following kernels?

- linear: $k(x_i, x_j) = \alpha x_i^T x_j$.
- polynomial: $k(x_i, x_j) = \alpha (x_i^T x_j + 1)^2$.
- RBF: $k(x_i, x_j) = \alpha_1 e^{-\alpha_2 \|x_i - x_j\|^2}$.
- periodic: $k(x_i, x_j) = \alpha_1 \exp\{-\alpha_2 \sin^2(\frac{x_i - x_j}{2})\}$.

α are the kernel parameters.

(e) $\hat{\sigma}_*^2$ is the variance of the predictive distribution of the function value f_*, which corresponds to the uncertainty of the prediction (high variance means high uncertainty). Under what conditions will the variance be large? When will the variance be small?

.....

Problem 10.13 Kernel perceptron

For a training set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, the Perceptron algorithm can be implemented as follows.

Algorithm 1 Perceptron algorithm

```
set w = 0, b = 0, R = max_i ||x_i||
repeat
  for i = 1, ..., n do
    if y_i(w^T x_i + b) ≤ 0 then
      set w ← w + η y_i x_i
      set b ← b + η y_i R^2
    end if
  end for
until there are no classification errors
```

The final classifier is $y_* = \text{sign}(w^T x_* + b)$.

(a) Is the learning rate η relevant? Why?

(b) Show that w learned by the perceptron algorithm must take the form $w = \sum_{i=1}^{n} \alpha_i y_i x_i$, where $\alpha_i \geq 0$, $\forall i$.

(c) Using (b), show that an equivalent Perceptron algorithm (the dual Perceptron) is:

Algorithm 2 Perceptron algorithm (dual)

```
set α = 0, b = 0, R = max_i ||x_i||
repeat
  for i = 1, ..., n do
    if y_i(Σ_{j=1}^{n} α_j y_j x_j^T x_i + b) ≤ 0 then
      set α_i ← α_i + 1
      set b ← b + y_i R^2
    end if
  end for
until there are no classification errors
```

(d) Can you give an interpretation to the parameters α_i? Which among the samples x_i are the hardest to classify?

(e) Apply the kernel trick the the dual perceptron algorithm to obtain the *kernel perceptron algorithm*. What is the kernelized decision function?

.....

Problem 10.14 Kernel k-means

In this problem we will kernelize the k-means algorithm.

Consider the original k-means algorithm. Given a set of points $X = \{x_1, \cdots, x_n\}$, where $x_i \in \mathbb{R}^d$, the goal is to assign a cluster label to each point, $y_i \in \{1, \cdots, K\}$, where K is the number of clusters. The k-means algorithm calculates the cluster centers μ_j using the current assignment to each cluster (K is assumed to be known). In each iteration, K-means performs the following two steps:

$$\text{Cluster Assignment}: \quad z_{ij} = \begin{cases} 1, & j = \text{argmin}_{k \in \{1,\cdots,K\}} \|x_i - \mu_k\|^2 \\ 0, & \text{otherwise.} \end{cases} \tag{10.37}$$

$$\text{Estimate Center}: \quad \mu_j = \frac{\sum_{i=1}^n z_{ij} x_i}{\sum_{i=1}^n z_{ij}}. \tag{10.38}$$

The cluster label for point x_i is the label of the closest cluster center, $y_i = \text{argmax}_j \, z_{ij}$.

The disadvantage of k-means is that it only works when the clusters can be separated by a hyperplane. This is a consequence of using the Euclidean distance to measure the distance to the cluster center. To apply the kernel trick, we need to rewrite k-means so that it only uses inner-products between the data points x_i.

(a) Show that the squared distance from x to the cluster center μ_k can be written only using inner-products as

$$d(x, \mu_k) = \|x - \mu_k\|^2 = x^T x - 2\frac{1}{N_k}\sum_{l=1}^n z_{lk} x^T x_l + \frac{1}{N_k^2}\sum_{l=1}^n \sum_{m=1}^n z_{lk} z_{mk} x_l^T x_m \tag{10.39}$$

where $N_k = \sum_{l=1}^n z_{mk}$ is the number of points in cluster k.

(b) Apply the kernel trick to (10.39) to obtain the kernelized square distance

$$d(x, \mu_k) = \|x - \mu_k\|^2 = k(x,x) - 2\frac{1}{N_k}\sum_{l=1}^n z_{ik} k(x, x_l) + \frac{1}{N_k^2}\sum_{l=1}^n \sum_{m=1}^n z_{lk} z_{mk} k(x_l, x_m). \tag{10.40}$$

(c) What is the interpretation of the distance in (10.40) when $k(x, x')$ is the Gaussian kernel? When will the distance be small?

(d) Show that the kernel k-means algorithm is:

$$\text{Calculate Distances}: \quad d(x_i, \mu_k) = k(x_i, x_i) - 2\frac{1}{N_k}\sum_{l=1}^n z_{ik} k(x_i, x_l) + \frac{1}{N_k^2}\sum_{l=1}^n \sum_{m=1}^n z_{lk} z_{mk} k(x_l, x_m), \forall i \tag{10.41}$$

$$\text{Cluster Assignment}: \quad z_{ij} = \begin{cases} 1, & j = \text{argmin}_{k \in \{1,\cdots,K\}} \, d(x_i, \mu_k) \\ 0, & \text{otherwise.} \end{cases} \tag{10.42}$$

A weighted version of Kernel k-means has interesting connection to many spectral clustering and graph clustering methods (see [I.S. Dhillon, Y. Guan, B. Kulis, "Kernel k-means: spectral clustering and normalized cuts." Proc. ACM SIGKDD, 2004]).

.....

Problem 10.15 Kernel discriminant analysis

In this problem we will derive a kernel version of Fisher's linear discriminant (FLD, see Problem 7.6) Let $X_1 = [x_1, \cdots, x_{n_1}]$ and $X_2 = [x_1, \cdots, x_{n_2}]$ be the matrix of feature vectors from class 1 and class 2, and n_1 and n_2 are the number of feature vectors for class 1 and class 2, respectively. The class mean and scatter matrix are given by

$$\mu_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i, \quad S_j = \sum_{i=1}^{n_j} (x_i - \mu_j)(x_i - \mu_j)^T. \tag{10.43}$$

FLD finds the optimal projection that maximizes the ratio of the "between-class" scatter and the "within-class" scatter,

$$w^* = \operatorname*{argmax}_w J(w), \quad J(w) = \frac{w^T S_B w}{w^T S_W w}, \tag{10.44}$$

where S_B and S_W are the between- and within-class scatter matrices,

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T, \quad S_W = S_1 + S_2. \tag{10.45}$$

The optimal projection is $w = S_W^{-1}(\mu_1 - \mu_2)$.

Let $X = [X_1, X_2]$ be all the data. To do the kernelization, we will first assume that the optimal w can be written as a linear combination of the data points,

$$w = \sum_{i=1}^{n} \alpha_i x_i = X\alpha, \tag{10.46}$$

where $\alpha = [\alpha_1, \cdots, \alpha_n]$ are the weights.

(a) Why is the assumption in (10.46) valid?

(b) Show that the class mean can be written as

$$\mu_j = \frac{1}{n_j} X_j \mathbf{1}, \tag{10.47}$$

and the scatter matrix as

$$S_j = X_j (I - \tfrac{1}{n_j} \mathbf{1}\mathbf{1}^T) X_j^T, \tag{10.48}$$

where $\mathbf{1}$ is a vector of ones.

(c) Using (b), show the mean and scatter matrices when multiplied by w can be written in terms of α,

$$w^T \mu_j = \alpha^T \hat{\mu}_j, \quad w^T S_j w = \alpha^T \hat{S}_j \alpha \tag{10.49}$$

where

$$\hat{\mu}_j = \frac{1}{n_j} X^T X_j \mathbf{1}, \quad \hat{S}_j = X^T X_j (I - \tfrac{1}{n_j} \mathbf{1}\mathbf{1}^T) X_j^T X \tag{10.50}$$

(d) Apply the kernel trick to $\hat{\mu}_j$ and $\hat{S}_j$ to obtain expressions,

$$\hat{\mu}_j = \frac{1}{n_j} K_j \mathbf{1}, \quad \hat{S}_j = K_j (I - \tfrac{1}{n_j} \mathbf{1}\mathbf{1}^T) K_j, \tag{10.51}$$

where K_j is the kernel matrix between X and X_j, i.e.,

$$[K_j]_{i,i'} = k(x_i, x_{i'}), \quad x_i \in X, \; x_i \in X_j. \tag{10.52}$$

(e) Show that the kernelized version of (10.44) is

$$\alpha^* = \underset{\alpha}{\operatorname{argmax}}\, J(\alpha), \quad J(\alpha) = \frac{\alpha^T \hat{S}_B \alpha}{\alpha^T \hat{S}_W \alpha}, \tag{10.53}$$

where

$$\hat{S}_B = (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T, \quad \hat{S}_W = \hat{S}_1 + \hat{S}_2. \tag{10.54}$$

(f) Show that the optimal α^* is given by

$$\alpha^* = \hat{S}_W^{-1}(\hat{\mu}_2 - \hat{\mu}_1). \tag{10.55}$$

(Hint: similar to the original problem, only the direction of α matters, and not the magnitude.

(g) Finally, show that the kernelized projection of x is

$$z = \sum_{i=1}^{n} \alpha_i k(x, x_i). \tag{10.56}$$