EE6550 Machine Learning

Lecture Four – Kernel Methods

Chung-Chin Lu
Department of Electrical Engineering
National Tsing Hua University

March 20, 2017

1

Motivation

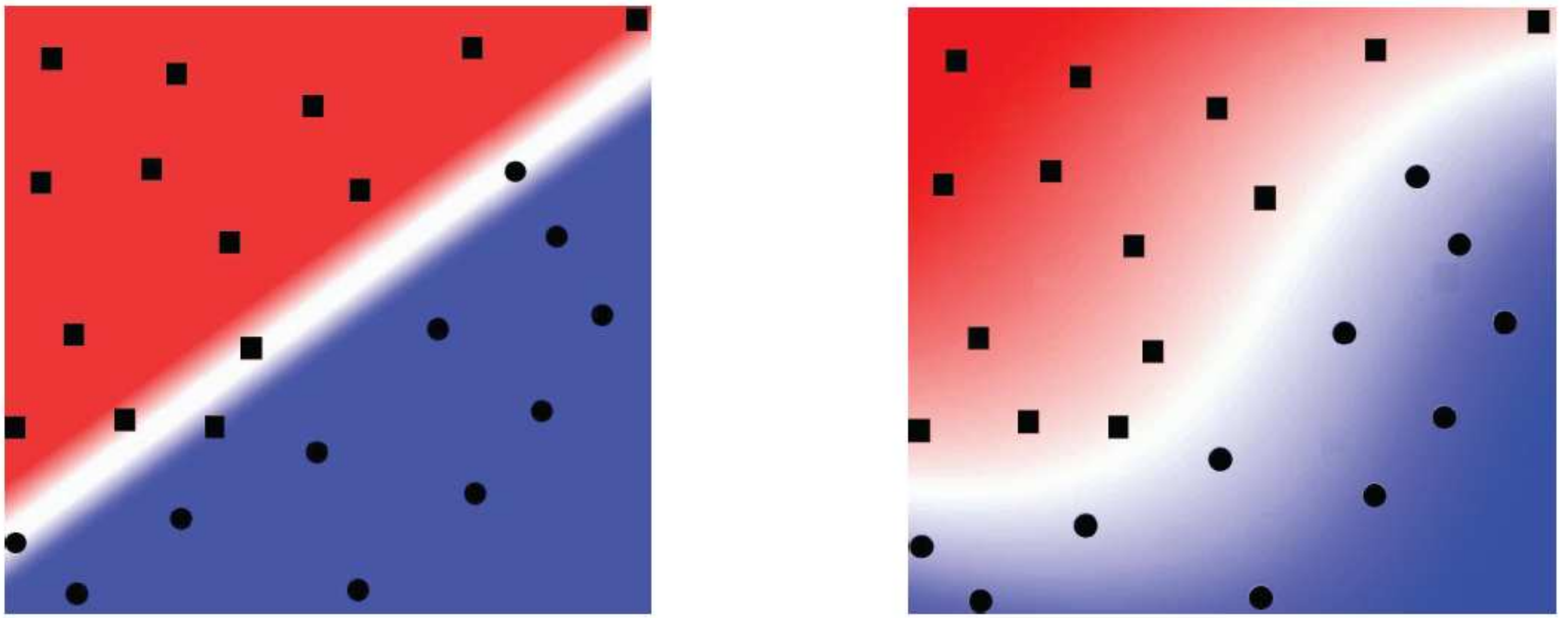
- Searching for large-margin separating hyperplanes in a very high-dimensional space.
  - Flexible selection of more complex features.
- Efficient computation of inner products in high dimension.
- Nonlinear decision boundary.
- Learning with non-vectorial inputs.

The Contents of This Lecture

- Positive definite symmetric (PDS) kernels
- Closure properties of PDS kernels
- Reproducing kernel Hilbert space (RKHS)
- SVMs with PDS kernels
- Conditionally negative definite symmetric (CNDS) kernels
- Sequence Kernels

Nonlinear Separation

- In most practical problems, perfect linear separation is usually impossible.
- Perfect nonlinear separation may be realized by a nonlinear mapping $\Phi : \mathscr{I} \to \mathscr{F}$ from the input space $\mathscr{I}$ to a high dimensional feature space $\mathscr{F}$.
- Margin-based bound gives a generalization guarantee which is independent of $\dim(\mathscr{F})$ but depends only on the confidence margin ρ and the sample size m.



(a) (b)

(a) No hyperplane can separate the two populations.

(b) A nonlinear mapping can be used instead.

Kernel Methods

- $\mathscr{I}$: the input space of all possible items, associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$.
- $\mathscr{F} = \mathbb{H}$: a chosen feature space, often a very high dimensional (or even infinite-dimensional) Hilbert space.
  - A Hilbert space is a complete inner product space.
- $\Phi : \mathscr{I} \to \mathscr{F}$: a feature mapping from the input space $\mathscr{I}$ to the feature space $\mathscr{F}$.
- $\langle \cdot, \cdot \rangle$: the inner product associated with the Hilbert space $\mathscr{F} = \mathbb{H}$ whose computation has very high cost if not impossible.
- Idea: using a kernel $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$ on the input space $\mathscr{I}$, defined as:

$$\forall\, \omega, \omega' \in \mathscr{I}, \quad K(\omega, \omega') \triangleq \langle \Phi(\omega), \Phi(\omega') \rangle.$$

- Benefits: efficiency and flexibility.
  - Efficiency: $K(\omega, \omega')$ is often more efficient to compute than $\Phi(\omega)$ and the inner product in $\mathbb{H}$.
  - Flexibility: K can be chosen arbitrarily without explicitly defining the feature space $\mathscr{F}$ and the feature mapping Φ as long as their existence is guaranteed (by the PDS condition or Mercer's condition).

Symmetric Positive Semi-Definite (SPSD) Matrices

An $m \times m$ real matrix $B = [b_{ij}]$ is called symmetric positive semi-definite (SPSD) if it is symmetric and one of the following two equivalent conditions holds:

1. all eigenvalues of B are non-negative;
2. for any m-tuple $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T \in \mathbb{R}^m$,

$$\mathbf{x}^T B \mathbf{x} = \sum_{i,j=1}^{m} x_i b_{ij} x_j \geq 0.$$

A Decomposition of an SPSD Matrix

- $\mathbf{B}$: an $m \times m$ SPSD matrix.
- $\lambda_i, 1 \leq i \leq m$: non-negative eigenvalues of $\mathbf{B}$.
- $\mathbf{v}_i, 1 \leq i \leq m$: orthonormal eigenvectors of $\mathbf{B}$ corresponding to eigenvalues λ_i respectively, $\mathbf{B}\mathbf{v}_i = \lambda_i \mathbf{v}_i,\ 1 \leq i \leq m$.
  - $\{\mathbf{v}_i, 1 \leq i \leq m\}$ is an orthonormal eigenbasis of $\mathbf{B}$ for $\mathbb{R}^m$.
- $\mathbf{Q} = [\mathbf{v}_1 \cdots \mathbf{v}_m]$: an $m \times m$ orthogonal matrix.
- $\mathbf{D} = \mathrm{diag}(\lambda_1, \cdots, \lambda_m)$: a diagonal $m \times m$ matrix with λ_i as diagonal entries.
- Since $\mathbf{BQ} = \mathbf{QD}$, we have

$$\mathbf{B} = \mathbf{QDQ}^T = (\mathbf{Q}\sqrt{\mathbf{D}})(\mathbf{Q}\sqrt{\mathbf{D}})^T = \mathbf{AA}^T,$$

where $\mathbf{A} = \mathbf{Q}\sqrt{\mathbf{D}}$.

Positive Definite Symmetric (PDS) Kernels

A kernel $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$ over the input space $\mathscr{I}$ is called positive definite symmetric if for any m-tuple $(\omega_1, \omega_2, \ldots, \omega_m)$ over $\mathscr{I}$, the $m \times m$ matrix $\mathbf{K} = [K(\omega_i, \omega_j)]$ is symmetric positive semi-definite (SPSD).

- If $S = (\omega_1, \omega_2, \ldots, \omega_m)$ is a sample of size m drawn i.i.d. from the input space $\mathscr{I}$ according to an unknown distribution P, the $m \times m$ matrix $\mathbf{K} = [K(\omega_i, \omega_j)]$ is called the kernel matrix or the Gram matrix associated to the kernel K and the sample S.

Kernels Defined by Inner Products Are PDS

Let

- $\mathscr{I}$: the input space of all possible items, associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$.
- $\mathbb{H}$: a Hilbert space, which is chosen as the feature space.
- $\Phi : \mathscr{I} \to \mathbb{H}$: a feature mapping from the input space to the feature space.
- $\langle \cdot, \cdot \rangle$: the inner product associated with the Hilbert space $\mathbb{H}$.

The kernel $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$ over the input space $\mathscr{I}$, defined as

$$\forall\, \omega, \omega' \in \mathscr{I}, \quad K(\omega, \omega') \triangleq \langle \Phi(\omega), \Phi(\omega') \rangle,$$

is positive definite symmetric (PDS).

Proof. Let

- $\mathbf{K} = [K(\omega_i, \omega_j)]$: the $m \times m$ real matrix associated with an m-tuple $(\omega_1, \omega_2, \ldots, \omega_m)$ over the input space $\mathscr{I}$;
- $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$: an m-tuple over $\mathbb{R}$.

Since the inner product is symmetric, we have

$$K(\omega_j, \omega_i) = \langle \Phi(\omega_j), \Phi(\omega_i) \rangle = \langle \Phi(\omega_i), \Phi(\omega_j) \rangle = K(\omega_i, \omega_j),$$

which shows that $\mathbf{K}$ is symmetric.

Also

$$\begin{aligned}
\mathbf{x}^T\mathbf{K}\mathbf{x} &= \sum_{i,j=1}^{m} x_i K(\omega_i, \omega_j) x_j \\
&= \sum_{i,j=1}^{m} x_i \langle \Phi(\omega_i), \Phi(\omega_j) \rangle x_j \\
&= \langle \sum_{i=1}^{m} x_i \Phi(\omega_i), \sum_{j=1}^{m} x_j \Phi(\omega_j) \rangle \geq 0,
\end{aligned}$$

by the positivity of inner product. Thus $\mathbf{K}$ is symmetric positive semi-definite and then K is positive definite symmetric. $\square$

Example 5.1: Polynomial Kernels

For any real constant c, a polynomial kernel of degree $d \geq 1$ is the kernel K over an input space $\mathscr{I} \subseteq \mathbb{R}^N$ defined as:
$\forall\ \mathbf{x} = (x_1, x_2, \ldots, x_N), \mathbf{x}' = (x'_1, x'_2, \ldots, x'_N) \in \mathscr{I}$,

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}') &\triangleq (c^2 + \mathbf{x} \cdot \mathbf{x}')^d = \left(c^2 + \sum_{i=1}^{N} x_i x'_i\right)^d \\ &= \sum_{\substack{d_0 + d_1 + \cdots + d_N = d \\ d_i \geq 0, 0 \leq i \leq N}} \frac{d!}{d_0! d_1! \cdots d_N!} (c^2)^{d_0} (x_1 x'_1)^{d_1} \cdots (x_N x'_N)^{d_N}. \end{aligned}$$

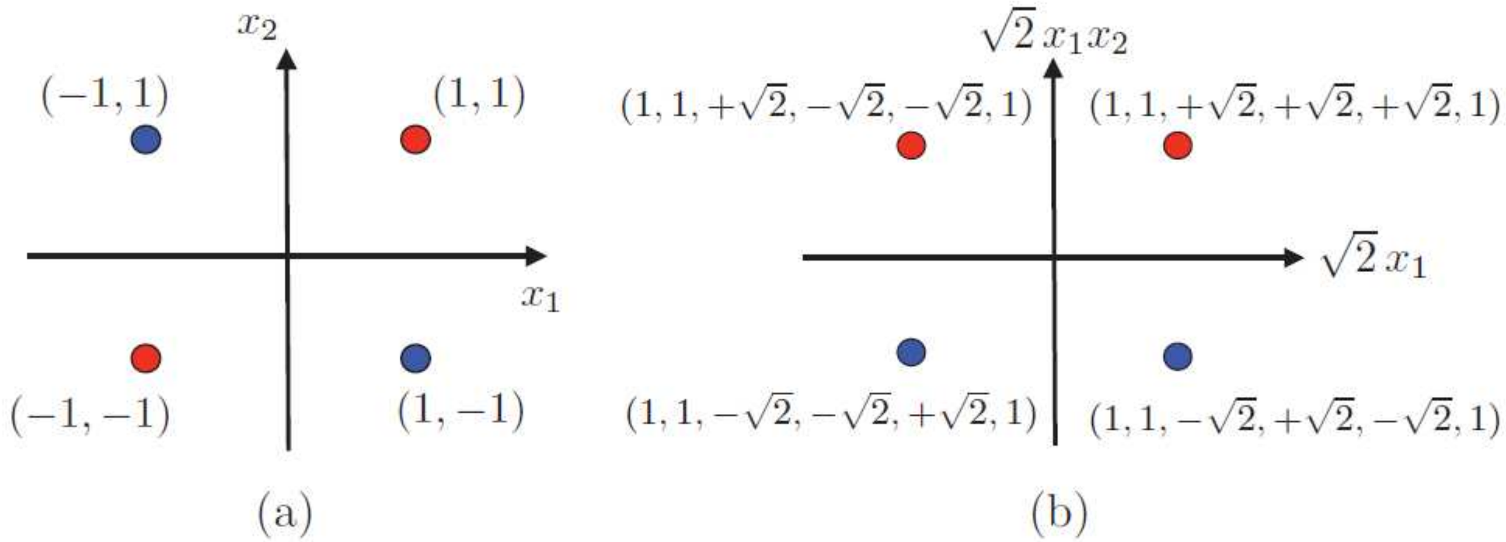
- There are $\begin{pmatrix} d + N \\ d \end{pmatrix}$ terms.

The Feature Space and Feature Mapping Associated to a Polynomial Kernel of Degree d

- $\mathscr{F} = \mathbb{R}^{\binom{d+N}{d}}$: the feature space, which is the Euclidean space of dimension $\binom{d+N}{d}$.
- $\Phi : \mathscr{I} \to \mathscr{F}$: the feature mapping defined as:

$$\Phi(\mathbf{x}) = \Big(\sqrt{\frac{d!}{d_0!d_1!\cdots d_N!}}c^{d_0}x_1^{d_1}\cdots x_N^{d_N}\Big)_{\substack{d_0+d_1+\cdots+d_N=d\\ d_i\geq 0, 0\leq i\leq N}}$$

- $K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle = \sum_{\substack{d_0+d_1+\cdots+d_N=d\\ d_i\geq 0, 0\leq i\leq N}} \frac{d!}{d_0!d_1!\cdots d_N!}(c^2)^{d_0}(x_1x_1')^{d_1}\cdots(x_Nx_N')^{d_N}$.
- K is PDS.



(a) XOR problem linearly nonseparable in the input space.

(b) Perfectly linearly separable using 2nd-degree polynomial kernel.

The Contents of This Lecture

- Positive definite symmetric (PDS) kernels
- Closure properties of PDS kernels
- Reproducing kernel Hilbert space (RKHS)
- SVMs with PDS kernels
- Conditionally negative definite symmetric (CNDS) kernels
- Sequence kernels

Cauchy-Schwarz Inequality for PDS Kernels

Lemma 5.1: Let

- K: a PDS kernel over an input space $\mathscr{I}$.

Then, for any $\omega, \omega' \in \mathscr{I}$,

$$K(\omega, \omega')^2 \leq K(\omega, \omega)K(\omega', \omega').$$

Proof. Consider the 2×2 matrix $\mathbf{K} = \begin{bmatrix} K(\omega, \omega) & K(\omega, \omega') \\ K(\omega', \omega) & K(\omega', \omega') \end{bmatrix}$.

Since K is PDS, $\mathbf{K}$ is SPSD and has non-negative eigenvalues and then

$$\det(\mathbf{K}) = K(\omega, \omega)K(\omega', \omega') - K(\omega, \omega')K(\omega', \omega) \geq 0.$$

By symmetry of K, we have $K(\omega, \omega') = K(\omega', \omega)$ and the inequality holds. $\square$

Normalized Kernel

Let

- $K' : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$: a kernel over the input space $\mathscr{I}$ such that $K(\omega, \omega) \geq 0$ for all $\omega \in \mathscr{I}$.

The normalized kernel K associated to K' is defined as:
$\forall\, \omega, \omega' \in \mathscr{I}$,

$$K(\omega, \omega') \triangleq \begin{cases} 0, & \text{if } K'(\omega, \omega) = 0 \text{ or } K'(\omega', \omega') = 0, \\ \frac{K'(\omega,\omega')}{\sqrt{K'(\omega,\omega)K'(\omega',\omega')}}, & \text{otherwise.} \end{cases}$$

- For a normalized kernel K, $K(\omega, \omega) = 1$ for all $\omega \in \mathscr{I}$ such that $K(\omega, \omega) \neq 0$.
- It is suggestive to know that for any PDS kernel K', if either $K'(\omega, \omega) = 0$ or $K'(\omega', \omega') = 0$, then $K'(\omega_i, \omega_j) = K'(\omega_j, \omega_i) = 0$ by Cauchy-Schwarz inequality.

Normalized PDS Kernels

Lemma 5.2: Let

- K': a PDS kernel.

Then the normalized kernel K associated to K' is also PDS.

Proof. Since K' is symmetric, K is also symmetric. Let

- $\mathbf{K} = [K(\omega_i, \omega_j)]$: the $m \times m$ real matrix associated with an m-tuple $(\omega_1, \omega_2, \ldots, \omega_m)$ over the input space $\mathscr{I}$;
- $I = \{i \in [1, m] : K'(\omega_i, \omega_i) = 0\}$;
- $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$: an m-tuple over $\mathbb{R}$.

By definition, $\forall\ i \in I,\ j \in [1, m]$,

$$K(\omega_i, \omega_j) = K(\omega_j, \omega_i) = 0.$$

Now we have

$$
\begin{aligned}
\mathbf{x}^T\mathbf{K}\mathbf{x} &= \sum_{i,j=1}^{m} x_i K(\omega_i, \omega_j) x_j \\
&= \sum_{i,j \notin I} x_i K(\omega_i, \omega_j) x_j \\
&= \sum_{i,j \notin I} \frac{x_i}{\sqrt{K'(\omega_i, \omega_i)}} K'(\omega_i, \omega_j) \frac{x_j}{\sqrt{K'(\omega_j, \omega_j)}} \\
&= \sum_{i,j=1}^{m} y_i K'(\omega_i, \omega_j) y_j \geq 0,
\end{aligned}
$$

where $y_i = 0$ if $i \in I$ and $y_i = \frac{x_i}{\sqrt{K'(\omega_i, \omega_i)}}$ if $i \notin I$. Thus $\mathbf{K}$ is symmetric positive semi-definite and then K is positive definite symmetric. $\square$

How to Combine PDS Kernels to Form New PDS Kernels?

Possible combinations are:

- Scalar multiplication. Let K be a kernel over an input space $\mathscr{I}$. The scalar multiplication aK of K by a scalar a is the kernel over $\mathscr{I}$ defined by: for all $\omega, \omega' \in \mathscr{I}$,

$$(aK)(\omega, \omega') = aK(\omega, \omega').$$

- Sum and product. Let K_1, K_2 be two kernels over an input space $\mathscr{I}$. For all $\omega, \omega' \in \mathscr{I}$,

$$\begin{aligned} \text{Sum}: (K_1 + K_2)(\omega, \omega') &\triangleq K_1(\omega, \omega') + K_2(\omega, \omega'), \\ \text{Product}: (K_1 K_2)(\omega, \omega') &\triangleq K_1(\omega, \omega') K_2(\omega, \omega'). \end{aligned}$$

- Tensor product. Let K_1 and K_2 be two kernels over input spaces $\mathscr{I}$ and $\mathscr{I}'$ respectively. The tensor product $K_1 \otimes K_2$ is a kernel over $\mathscr{I} \times \mathscr{I}'$ defined as: for all $(\omega, \varpi), (\omega', \varpi') \in \mathscr{I} \times \mathscr{I}'$,

$$K_1 \otimes K_2((\omega, \varpi), (\omega', \varpi')) \triangleq K_1(\omega, \omega') K_2(\varpi, \varpi').$$

- Pointwise limit. Let $K_1, K_2, \ldots, K_n, \ldots$ be a sequence of kernels over an input space $\mathscr{I}$ such that for each ordered pair (ω, ω') over $\mathscr{I}$, the limit $\lim_{n\to\infty} K_n(\omega, \omega')$ exists. The limit $K = \lim_{n\to\infty} K_n$ of the sequence $\{K_n\}$ is the kernel over $\mathscr{I}$, defined as: for all $\omega, \omega' \in \mathscr{I}$,

$$K(\omega, \omega') \triangleq \lim_{n\to\infty} K_n(\omega, \omega').$$

- Composition with a power series. Let $\sum_{n=0}^{\infty} a_n x^n$ be a power series with radius of convergence $\rho > 0$ and K a kernel taking values in $(-\rho, +\rho)$. The power series $\sum_{n=0}^{\infty} a_n K^n$ of K is the kernel over $\mathscr{I}$, defined as: for all $\omega, \omega' \in \mathscr{I}$,

$$\left(\sum_{n=0}^{\infty} a_n K^n\right)(\omega, \omega') \triangleq \sum_{n=0}^{\infty} a_n K^n(\omega, \omega').$$

Closure Properties of PDS Kernels

Theorem 5.3: PDS kernels are closed under scalar multiplication by a scalar $a \geq 0$, sum, product, tensor product, pointwise limit, and composition with a power series $\sum_{n=0}^{\infty} a_n x^n$ with $a_n \geq 0$ for all n.

Proof.

- Scalar multiplication.
  - Since K is symmetric, aK is also symmetric.
  - Let $\mathbf{K}$ be an $m \times m$ matrix associated with an m-tuple $(\omega_1, \omega_2, \ldots, \omega_m)$ over the input space $\mathscr{I}$ for the PDS kernel K. It is SPSD.
  - Then $a\mathbf{K}$ is the $m \times m$ matrix associated with the m-tuple $(\omega_1, \omega_2, \ldots, \omega_m)$ over the input space $\mathscr{I}$ for the kernel aK.
  - Since $a \geq 0$ and $\mathbf{K}$ is SPSD, $a\mathbf{K}$ is also SPSD and then aK is PDS.

25

- Sum and product.
  - Since K_1 and K_2 are symmetric, their sum $K_1 + K_2$ and product K_1K_2 are also symmetric.
  - Let $\mathbf{K}_1, \mathbf{K}_2$ be two $m \times m$ matrices associated with an m-tuple $(\omega_1, \omega_2, \ldots, \omega_m)$ over the input space $\mathscr{I}$ for two PDS kernels K_1 and K_2 respectively. They are SPSD.
  - Then $\mathbf{K}_1 + \mathbf{K}_2$ is the $m \times m$ matrix associated with the m-tuple $(\omega_1, \omega_2, \ldots, \omega_m)$ over the input space $\mathscr{I}$ for the sum kernel $K_1 + K_2$.
  - Let $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$ be an m-tuple over $\mathbb{R}$.
  - Since $\mathbf{x}^T\mathbf{K}_1\mathbf{x} \geq 0$ and $\mathbf{x}^T\mathbf{K}_2\mathbf{x} \geq 0$, we have $\mathbf{x}^T(\mathbf{K}_1 + \mathbf{K}_2)\mathbf{x} \geq 0$ so that $\mathbf{K}_1 + \mathbf{K}_2$ is SPSD and then the sum $K_1 + K_2$ is PDS.

– Since $\mathbf{K}_1$ is SPSD, there exists an $m \times m$ matrix $\mathbf{A} = [a_{ij}]$ such that $\mathbf{K}_1 = \mathbf{A}\mathbf{A}^T$, i.e., $K_1(\omega_i, \omega_j) = \sum_{k=1}^{m} a_{ik} a_{kj}$.

– The matrix $\mathbf{K} \triangleq [K_1(\omega_i, \omega_j) K_2(\omega_i, \omega_j)]$ is the $m \times m$ matrix associated with the m-tuple $(\omega_1, \omega_2, \ldots, \omega_m)$ over the input space $\mathscr{I}$ for the product kernel $K_1 K_2$.

– Now we have

$$\begin{aligned}
\mathbf{x}^T \mathbf{K} \mathbf{x} &= \sum_{i,j=1}^{m} x_i K_1(\omega_i, \omega_j) K_2(\omega_i, \omega_j) x_j \\
&= \sum_{i,j=1}^{m} x_i \sum_{k=1}^{m} a_{ik} a_{kj} K_2(\omega_i, \omega_j) x_j \\
&= \sum_{k=1}^{m} \sum_{i,j=1}^{m} (x_i a_{ik}) K_2(\omega_i, \omega_j)(x_j a_{kj}) \geq 0,
\end{aligned}$$

since $\mathbf{K}_2$ is SPSD, which says that $\mathbf{K}$ is SPSD and then $K_1 K_2$ is PDS.

- Tensor product.
  - Define two kernels $\tilde{K}_1$ and $\tilde{K}_2$ over the the Cartesian product $\mathscr{I} \times \mathscr{I}'$ of input spaces $\mathscr{I}$ and $\mathscr{I}'$: for all $(\omega, \varpi), (\omega', \varpi') \in \mathscr{I} \times \mathscr{I}'$,

$$\begin{aligned} \tilde{K}_1((\omega, \varpi), (\omega', \varpi')) &\triangleq K_1(\omega, \omega'), \\ \tilde{K}_2((\omega, \varpi), (\omega', \varpi')) &\triangleq K_2(\varpi, \varpi'). \end{aligned}$$

  - Since K_1 and K_2 are symmetric, $\tilde{K}_1$ and $\tilde{K}_2$ are also symmetric.
  - Let $\tilde{\mathbf{K}}_1, \tilde{\mathbf{K}}_2$ be two $m \times m$ matrices associated with an m-tuple $((\omega_1, \varpi_1), (\omega_2, \varpi_2), \ldots, (\omega_m, \varpi_m))$ over the Cartesian product input space $\mathscr{I} \times \mathscr{I}'$ for the two kernels $\tilde{K}_1$ and $\tilde{K}_2$ respectively.
  - Since $\tilde{\mathbf{K}}_1 = [\tilde{K}_1((\omega_i, \varpi_i), (\omega_j, \varpi_j))] = [K_1(\omega_i, \omega_j)]$, $\tilde{\mathbf{K}}_1$ is SPSD and then $\tilde{K}_1$ is PDS.

- Similarly since $\tilde{\mathbf{K}}_2 = [\tilde{K}_2((\omega_i, \varpi_i), (\omega_j, \varpi_j))] = [K_2(\varpi_i, \varpi_j)]$, $\tilde{\mathbf{K}}_2$ is also SPSD and then $\tilde{K}_2$ is PDS.
- It can be seen that the tensor product $K_1 \otimes K_2$ of K_1 and K_2 is the product $\tilde{K}_1 \tilde{K}_2$ of $\tilde{K}_1$ and $\tilde{K}_2$.
- Since both $\tilde{K}_1$ and $\tilde{K}_2$ are PDS, the tensor product $K_1 \otimes K_2 = \tilde{K}_1 \tilde{K}_2$ is also PDS.

- Pointwise limit.
  - Let the limit $K = \lim_{n\to\infty} K_n$ of the sequence $\{K_n\}$ exist.
  - Since K_n's are symmetric, the limit K is also symmetric.
  - Let $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_n, \ldots$ be the sequence of $m \times m$ matrices associated with an m-tuple $(\omega_1, \omega_2, \ldots, \omega_m)$ over the input space $\mathscr{I}$ for a sequence $K_1, K_2, \ldots, K_n, \ldots$ of kernels respectively. They are SPSD.
  - The matrix $\mathbf{K} = [\lim_{n\to\infty} K_n(\omega_i, \omega_j)]$ is the $m \times m$ matrices associated with the m-tuple $(\omega_1, \omega_2, \ldots, \omega_m)$ over the input

space $\mathscr{I}$ for the limit kernel $K = \lim_{n\to\infty} K_n$.

- $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$: an m-tuple over $\mathbb{R}$.
- Now we have

$$\begin{aligned}\mathbf{x}^T\mathbf{K}\mathbf{x} &= \sum_{i,j=1}^{m} x_i \lim_{n\to\infty} K_n(\omega_i, \omega_j) x_j \\ &= \lim_{n\to\infty} \sum_{i,j=1}^{m} x_i K_n(\omega_i, \omega_j) x_j \geq 0,\end{aligned}$$

which says that $\mathbf{K}$ is SPSD and then the limit $K = \lim_{n\to\infty} K_n$ is PDS.

- Composition with a power series.
  - Since the kernel K is PDS, its powers K^i are also PDS for all $i \geq 0$.
  - Since $a_i \geq 0$, $a_i K^i$ are PDS for all $i \geq 0$.
  - The partial sums $\sum_{i=0}^{n} a_i K^i$ are PDS for all $n \geq 0$.
  - Since K takes values within the region of convergence of the power series $\sum_{n=0}^{\infty} a_n x^n$, the power series $\sum_{n=0}^{\infty} a_n K^n$, as the limit $\lim_{n\to\infty} \sum_{i=0}^{n} a_i K^i$ of partial sums, exists and is PDS. $\square$

Remarks

- Since the power series expansion $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ of the exponential function e^x has non-negative coefficients and infinite radius of convergence,

$$\exp(K(\omega, \omega')) \triangleq \sum_{n=0}^{\infty} \frac{K(\omega, \omega')^n}{n!}$$

  is a PDS kernel if K is a PDS kernel.

- $K'(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$: an inner product kernel over an input space $\mathscr{I}$ contained in a Hilbert space $\mathbb{H}$, which is PDS.

- $\left(\frac{K'}{\sigma^2}\right)(\mathbf{x}, \mathbf{x}') = \frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\sigma^2}$: a PDS kernel over $\mathscr{I} \subseteq \mathbb{H}$ for any $\sigma > 0$.

- $\exp\left(\frac{K'}{\sigma^2}\right)(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\sigma^2}\right)$: a PDS kernel over the input space $\mathscr{I} \subseteq \mathbb{H}$.

Example 5.2: Gaussian Kernels

For any constant $\sigma > 0$, a Gaussian kernel or radial basis function (RBF) is the kernel K over an input space $\mathscr{I} \subseteq \mathbb{R}^N$ defined as:
$\forall\ \mathbf{x} = (x_1, x_2, \ldots, x_N), \mathbf{x}' = (x'_1, x'_2, \ldots, x'_N) \in \mathscr{I}$,

$$K(\mathbf{x}, \mathbf{x}') \quad \triangleq \quad \exp\{\frac{-\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\}.$$

- A Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp\{\frac{-\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\}$ is the normalization of the PDS kernel $K'(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{\mathbf{x}\cdot\mathbf{x}'}{\sigma^2}\right)$ since

$$\begin{aligned}
\frac{K'(\mathbf{x}, \mathbf{x}')}{\sqrt{K'(\mathbf{x}, \mathbf{x})K'(\mathbf{x}', \mathbf{x}')}} &= \exp\left(\frac{-\|\mathbf{x}\|^2 - \|\mathbf{x}'\|^2 + 2\mathbf{x}\cdot\mathbf{x}'}{2\sigma^2}\right) \\
&= \exp\{\frac{-\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\}.
\end{aligned}$$

- Gaussian kernels are PDS.

The Contents of This Lecture

- Positive definite symmetric (PDS) kernels
- Closure properties of PDS kernels
- Reproducing kernel Hilbert space (RKHS)
- SVMs with PDS kernels
- Conditionally negative definite symmetric (CNDS) kernels
- Sequence kernels

Reproducing Kernel Hilbert Space (RKHS)

Theorem 5.2: Let

- K: a PDS kernel over an input space $\mathscr{I}$.

Then, there exists a Hilbert space $\mathbb{H}$ and a feature mapping Φ from $\mathscr{I}$ to $\mathbb{H}$ such that:

$$\forall\, \omega, \omega' \in \mathscr{I}, \quad K(\omega, \omega') = \langle \Phi(\omega), \Phi(\omega') \rangle.$$

Furthermore, $\mathbb{H}$ has the following property known as the reproducing property:

$$\forall\, f \in \mathbb{H},\ \omega \in \mathscr{I}, \quad f(\omega) = \langle f, K(\omega, \cdot) \rangle = \langle f, \Phi(\omega) \rangle.$$

$\mathbb{H}$ is called a reproducing kernel Hilbert space (RKHS) associated to the PDS kernel K.

Proof.

- For each $\omega \in \mathscr{I}$, define a real-valued function $\Phi(\omega) : \mathscr{I} \to \mathbb{R}$ over the input space $\mathscr{I}$ as follows:

$$\Phi(\omega)(\omega') \triangleq K(\omega, \omega'), \ \forall \, \omega' \in \mathscr{I}.$$

- $\mathbb{H}_0 = \mathrm{Span}\{\Phi(\omega) : \omega \in \mathscr{I}\}$: the set of linear combinations of finite number of functions $\Phi(\omega)$, $\omega \in \mathscr{I}$.
  - $\mathbb{H}_0$ is a vector space over $\mathbb{R}$.
- $\langle \cdot, \cdot \rangle$: a map from $\mathbb{H}_0 \times \mathbb{H}_0$ to $\mathbb{R}$, defined by: for all $f = \sum_i a_i \Phi(\omega_i)$, $g = \sum_j b_j \Phi(\omega'_j) \in \mathbb{H}_0$,

$$\langle f, g \rangle \triangleq \sum_{ij} a_i b_j K(\omega_i, \omega'_j) = \sum_j b_j f(\omega'_j) = \sum_i a_i g(\omega_i).$$

  - By definition, $\langle \cdot, \cdot \rangle$ is symmetric.
  - By the last two equalities, $\langle \cdot, \cdot \rangle$ is well-defined and bilinear.

– Also $\langle f, f\rangle = \sum_{ij} a_i a_j K(\omega_i, \omega_j) \geq 0$ since K is PDS.

– $\langle \cdot, \cdot \rangle$ is a positive semi-definite bilinear form on the vector space $\mathbb{H}_0$.

- $\langle \cdot, \cdot \rangle$: a PDS kernel over $\mathbb{H}_0$ since

$$\sum_{ij} a_i a_j \langle f_i, f_j\rangle = \langle \sum_i a_i f_i, \sum_j a_j f_j \rangle \geq 0, \ \forall\, f_i \in \mathbb{H}_0 \text{ and } \forall a_i \in \mathbb{R}.$$

- By Cauchy-Schwarz inequality, for any $f \in \mathbb{H}_0$ and $\omega \in \mathscr{I}$,

$$\langle f, \Phi(\omega)\rangle^2 \leq \langle f, f\rangle \langle \Phi(\omega), \Phi(\omega)\rangle.$$

- The reproducing property of $\langle \cdot, \cdot \rangle$: for any $f = \sum_i a_i \Phi(\omega_i) \in \mathbb{H}_0$ and $\omega \in \mathscr{I}$,

$$\forall\, \omega \in \mathscr{I}, \quad f(\omega) = \sum_i a_i \Phi(\omega_i)(\omega) = \sum_i a_i K(\omega_i, \omega) = \langle f, \Phi(\omega)\rangle.$$

- Thus we have $|f(\omega)|^2 \leq \langle f, f\rangle K(\omega, \omega)$.

- If $f \in \mathbb{H}_0$ is not the zero function, i.e., there is an $\omega \in \mathscr{I}$ such that $f(\omega) \neq 0$, then we have $\langle f, f\rangle K(\omega, \omega) > 0$ and then $\langle f, f\rangle > 0$. This shows that $\langle \cdot, \cdot \rangle$ is positive definite and then an inner product on $\mathbb{H}_0$.

- The inner product space $\mathbb{H}_0$ can be completed to form a Hilbert space $\mathbb{H}$ in which it is dense, following a standard construction.

- By the Cauchy-Schwarz inequality, for any $\omega \in \mathscr{I}$, the function $f \mapsto \langle f, \Phi(\omega)\rangle$ on $\mathbb{H}$ is Lipschitz,

$$\begin{aligned} & |\langle f_1, \Phi(\omega)\rangle - \langle f_2, \Phi(\omega)\rangle| = |\langle f_1 - f_2, \Phi(\omega)\rangle| \\ \leq \quad & \sqrt{\langle f_1 - f_2, f_1 - f_2\rangle}\sqrt{K(\omega, \omega)} = \sqrt{K(\omega, \omega)}\|f_1 - f_2\| \end{aligned}$$

and therefore continuous. Since $\mathbb{H}_0$ is dense in $\mathbb{H}$, the reproducing property also holds over $\mathbb{H}$. $\square$

Remarks

- The Hilbert space $\mathbb{H}$ defined in the proof of the theorem for a PDS kernel K is called the reproducing kernel Hilbert space (RKHS) associated to K.
- Any Hilbert space $\mathbb{H}$ such that there exists $\Phi : \mathscr{I} \to \mathbb{H}$ with $K(\omega, \omega') = \langle \Phi(\omega), \Phi(\omega') \rangle$ for all $\omega, \omega' \in \mathscr{I}$ is called a feature space associated to K and Φ is called a feature mapping.
- The feature spaces associated to K are in general not unique and may have different dimensions.
- In practice, when referring to the dimension of the feature space associated to K, we either refer to the dimension of the feature space based on a feature mapping described explicitly, or to that of the RKHS associated to K.

Remarks

- While one of the advantages of PDS kernels is an implicit definition of a feature mapping, in some instances, it may be desirable to define an explicit feature mapping based on a PDS kernel.
- This may be required to work in the primal problems for various optimization and computational reasons, to derive an approximation based on an explicit mapping, or as part of a theoretical analysis where an explicit mapping is more convenient

Empirical Kernel Maps Associated to a PDS Kernel

- $\mathscr{I}$: the input space of all possible items, associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$, where P is unknown.
- K: a PDS kernel over the input space $\mathscr{I}$.
- $S = (\omega_1, \omega_2, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from the input space $\mathscr{I}$ according to the distribution P.

The empirical kernel map Φ_S associated to a PDS kernel K under the sample S of size m is a mapping from $\mathscr{I}$ to $\mathbb{R}^m$: for all $\omega \in \mathscr{I}$,

$$\Phi_S(\omega) = \begin{bmatrix} K(\omega, \omega_1) \\ \vdots \\ K(\omega, \omega_m) \end{bmatrix}.$$

- $\mathbb{R}^m$: the empirical feature space under the sample S of size m.

- $\Phi_S(\omega)$ is the vector of the K-similarity measures of ω with each of the training points ω_i in the sample S.

Empirical Kernels K_S

The empirical kernel K_S associated to the PDS kernel K and the sample $S = (\omega_1, \omega_2, \ldots, \omega_m)$ of size m is defined by the empirical kernel map Φ_S from the input space $\mathscr{I}$ to the empirical feature space $\mathbb{R}^m$ as follows: for all $\omega, \omega' \in \mathscr{I}$,

$$K_S(\omega, \omega') \triangleq \Phi_S(\omega)^T \Phi_S(\omega') = \sum_{k=1}^{m} K(\omega, \omega_k) K(\omega_k, \omega').$$

- K_S is PDS.
- Since $\Phi_S(\omega)^T \Phi_S(\omega') = \sum_{k=1}^{m} K(\omega, \omega_k) K(\omega_k, \omega')$ may not be equal to $K(\omega, \omega')$, K_S is in general not equal to the original PDS kernel K.
- The kernel matrix $\mathbf{K}_S = [K_S(\omega_i, \omega_j)]$ associated to the

empirical kernel K_S and the sample S is

$$K_S(\omega_i, \omega_j) = \sum_{k=1}^{m} K(\omega_i, \omega_k) K(\omega_k, \omega_j) = (\mathbf{K}^2)_{ij},$$

where $\mathbf{K} = [K(\omega_i, \omega_j)]$ is the kernel matrix associated to the kernel K and the sample S, so that

$$\mathbf{K}_S = \mathbf{K}^2.$$

- To define a type of empirical kernels such that the kernel matrix associated to such an empirical kernel and the sample S is the same as the kernel matrix $\mathbf{K}$ associated to the kernel K and the sample S, we need pseudoinverse of $\mathbf{K}$.

Singular Values of a Rectangular Matrix

- $\mathbf{A}$: an $m \times n$ real matrix.
- $\mathbf{A}^T\mathbf{A}$: an $n \times n$ symmetric positive semi-definite matrix.
- $\lambda_i, 1 \leq i \leq n$: n non-negative eigenvalues of $\mathbf{A}^T\mathbf{A}$.
- $\mathbf{v}_i, 1 \leq i \leq n$: orthonormal eigenvectors of $\mathbf{A}^T\mathbf{A}$ corresponding to eigenvalues λ_i respectively,

$$\mathbf{A}^T\mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i,\ 1 \leq i \leq n.$$

  - $\{\mathbf{v}_i, 1 \leq i \leq n\}$ is an orthonormal eigenbasis of $\mathbf{A}^T\mathbf{A}$ in $\mathbb{R}^n$.
- $\sqrt{\lambda_i}, 1 \leq i \leq n$: singular values of $\mathbf{A}$.

The Action of $\mathbf{A}$ on the Orthonormal Eigenbasis $\{\mathbf{v}_i, 1 \le i \le n\}$ of $\mathbf{A}^T\mathbf{A}$

$$(\mathbf{A}\mathbf{v}_i)^T (\mathbf{A}\mathbf{v}_j) = \mathbf{v}_i^T \mathbf{A}^T \mathbf{A}\mathbf{v}_j = \lambda_j \mathbf{v}_i^T \mathbf{v}_j = \lambda_j \delta_{ij}.$$

- $\{\mathbf{A}\mathbf{v}_i, 1 \le i \le n\}$: orthogonal vectors in $\mathbb{R}^m$.
- $\|\mathbf{A}\mathbf{v}_i\|^2 = \lambda_i$.
- Number of non-zero λ_i = the rank of $\mathbf{A}$.

Singular Value Decomposition (SVD) of A

- r : the rank of $\mathbf{A}$.
- $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r}$: non-zero singular values of $\mathbf{A}$.
- $\{\mathbf{A}\mathbf{v}_1/\sqrt{\lambda_1}, \ldots, \mathbf{A}\mathbf{v}_r/\sqrt{\lambda_r}\}$: an orthonormal set in $\mathbb{R}^m$.
- $\{\mathbf{u}_j, 1 \leq j \leq m\}$: an orthonormal basis of $\mathbb{R}^m$ with

$$\mathbf{u}_j = \mathbf{A}\mathbf{v}_j/\sqrt{\lambda_j}, \forall\ 1 \leq j \leq r.$$

Since

$$\mathbf{A}\mathbf{v}_i = \begin{cases} \sqrt{\lambda_i}\mathbf{u}_i, & \text{if } 1 \leq i \leq r, \\ 0, & \text{if } r+1 \leq i \leq n, \end{cases}$$

we have

$$
\begin{aligned}
& \mathbf{A}[\mathbf{v}_1\mathbf{v}_2\cdots\mathbf{v}_r\mathbf{v}_{r+1}\cdots\mathbf{v}_n] \\
= \; & [\mathbf{u}_1\mathbf{u}_2\cdots\mathbf{u}_r\mathbf{u}_{r+1}\cdots\mathbf{u}_m]
\begin{bmatrix}
\sqrt{\lambda_1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & \sqrt{\lambda_2} & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sqrt{\lambda_r} & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 & \cdots & 0
\end{bmatrix}.
\end{aligned}
$$

Let $\mathbf{V} \triangleq [\mathbf{v}_1\mathbf{v}_2\cdots\mathbf{v}_n]$ and $\mathbf{U} \triangleq [\mathbf{u}_1\mathbf{u}_2\cdots\mathbf{u}_m]$, which are $n \times n$ and

$m \times m$ orthogonal matrices respectively. Let

$$\mathbf{\Sigma} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_r} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix},$$

which is a diagonal matrix. Then we have

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{U}_{\mathbf{A}}\mathbf{\Sigma}_{\mathbf{A}}\mathbf{V}_{\mathbf{A}}^T,$$

which is called the singular value decomposition of $\mathbf{A}$, where

- $\mathbf{\Sigma}_{\mathbf{A}} = \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_r})$ is an $r \times r$ diagonal matrix;

- $\mathbf{V}_{\mathbf{A}} = [\mathbf{v}_1 \mathbf{v}_2 \cdots \mathbf{v}_r]$ is an $n \times r$ matrix;
- $\mathbf{U}_{\mathbf{A}} = [\mathbf{u}_1 \mathbf{u}_2 \cdots \mathbf{u}_r]$ is an $m \times r$ matrix.

Remarks

- $\lambda_1, \lambda_2, \ldots, \lambda_r$ are all non-zero eigenvalues of the $m \times m$ SPSD matrix $\mathbf{A}\mathbf{A}^T$ and $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$ are corresponding eigenvectors respectively.

  Proof. For each $j \in [1, r]$, we have

  $$\mathbf{u}_j = \mathbf{A}\mathbf{v}_j / \sqrt{\lambda_j}$$

  and then

  $$\mathbf{A}\mathbf{A}^T\mathbf{u}_j = \mathbf{A}\mathbf{A}^T\mathbf{A}\mathbf{v}_j/\sqrt{\lambda_j} = \lambda_j \mathbf{A}\mathbf{v}_j/\sqrt{\lambda_j} = \lambda_j \mathbf{u}_j.$$

  Thus $\lambda_1, \lambda_2, \ldots, \lambda_r$ are non-zero eigenvalues of $\mathbf{A}\mathbf{A}^T$. If there were other non-zero eigenvalues of $\mathbf{A}\mathbf{A}^T$, then they must be non-zero eigenvalues of $\mathbf{A}^T\mathbf{A}$ by similar argument, which is a contradiction. Thus $\lambda_1, \lambda_2, \ldots, \lambda_r$ are all non-zero eigenvalues of $\mathbf{A}\mathbf{A}^T$. $\square$

- $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m$ are eigenvectors of $\mathbf{A}\mathbf{A}^T$ corresponding to eigenvalue 0.

  Proof. Since eigenvectors corresponding to distinct eigenvalues of a symmetric matrix are orthogonal, the eigenspace corresponding to the eigenvalue 0 of $\mathbf{A}\mathbf{A}^T$ is the orthogonal complement of the subspace spanned by eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$ corresponding to all non-zero eigenvalues. Since $\mathrm{Span}(\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m)$ is the orthogonal complement of $\mathrm{Span}(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r)$, $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m$ are eigenvectors of $\mathbf{A}\mathbf{A}^T$ corresponding to eigenvalue 0. $\square$

- An eigenvector $\mathbf{v}_i$ of $\mathbf{A}^T\mathbf{A}$ corresponding to eigenvalue λ_i is called a right-singular vector of $\mathbf{A}$ and the corresponding eigenvector $\mathbf{u}_i$ of $\mathbf{A}\mathbf{A}^T$ is called the left-singular vector of $\mathbf{A}$ corresponding to the right-singular vector $\mathbf{v}_i$.

- We have

$$\mathbf{A}^T\mathbf{u}_i = \begin{cases} \sqrt{\lambda_i}\mathbf{v}_i, & \text{if } \ 1 \leq i \leq r, \\ 0, & \text{if } \ r+1 \leq i \leq m, \end{cases}$$

- If $\mathbf{A}$ is symmetric, i.e., $\mathbf{A} = \mathbf{A}^T$, then $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{A}^2$ and the singular values of $\mathbf{A}$ are the absolute values of eigenvalues of $\mathbf{A}$. Any eigenvector $\mathbf{v}_i$ of $\mathbf{A}$ corresponding to an eigenvalues μ_i of $\mathbf{A}$ is a right-singular vector of $\mathbf{A}$ corresponding to the singular value $\sqrt{\lambda_i} = |\mu_i|$ of $\mathbf{A}$ and $\mathbf{u}_i = \operatorname{sgn}(\mu_i)\mathbf{v}_i$ is the left-singular vector of $\mathbf{A}$ corresponding to the right-singular vector $\mathbf{v}_i$. Thus an SVD of $\mathbf{A}$ is

$$
\begin{aligned}
\mathbf{A} &= [\mathrm{sgn}(\mu_1)\mathbf{v}_1 \cdots \mathrm{sgn}(\mu_r)\mathbf{v}_r \ \mathrm{sgn}(\mu_{r+1})\mathbf{v}_{r+1} \cdots \mathrm{sgn}(\mu_n)\mathbf{v}_n] \\
&\quad \begin{bmatrix} \mathrm{diag}(|\mu_1|, \ldots, |\mu_r|) & \mathbf{O}_{r\times(n-r)} \\ \mathbf{O}_{(n-r)\times r} & \mathbf{O}_{(n-r)\times(n-r)} \end{bmatrix} \left[\mathbf{v}_1 \cdots \mathbf{v}_r \ \mathbf{v}_{r+1} \cdots \mathbf{v}_n\right]^T \\
&= \left[\mathbf{v}_1 \cdots \mathbf{v}_r \ \mathbf{v}_{r+1} \cdots \mathbf{v}_n\right] \\
&\quad \begin{bmatrix} \mathrm{diag}(\mu_1, \ldots, \mu_r) & \mathbf{O}_{r\times(n-r)} \\ \mathbf{O}_{(n-r)\times r} & \mathbf{O}_{(n-r)\times(n-r)} \end{bmatrix} \left[\mathbf{v}_1 \cdots \mathbf{v}_r \ \mathbf{v}_{r+1} \cdots \mathbf{v}_n\right]^T \\
&= \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T,
\end{aligned}
$$

which is just a spectral decomposition of the symmetric matrix $\mathbf{A}$.

Moore-Penrose Pseudoinverse of a Rectangular Matrix

A (Moore-Penrose) pseudoinverse of an $m \times n$ real matrix $\mathbf{A}$ is an $n \times m$ real matrix $\mathbf{A}^+$ such that

1. $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$;
2. $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$;
3. $(\mathbf{A}\mathbf{A}^+)^T = \mathbf{A}\mathbf{A}^+$;
4. $(\mathbf{A}^+\mathbf{A})^T = \mathbf{A}^+\mathbf{A}$.

Uniqueness of Pseudoinverse

Let $\mathbf{A}^+$ and $\mathbf{B}^+$ be two pseudoinverses of $\mathbf{A}$. We first show that

$$\mathbf{A}\mathbf{A}^+ = \mathbf{A}\mathbf{B}^+ \quad \text{and} \quad \mathbf{A}^+\mathbf{A} = \mathbf{B}^+\mathbf{A}.$$

These are because

$$\begin{aligned}
\mathbf{A}\mathbf{A}^+ &= (\mathbf{A}\mathbf{A}^+)^T = (\mathbf{A}^+)^T\mathbf{A}^T = (\mathbf{A}^+)^T(\mathbf{A}\mathbf{B}^+\mathbf{A})^T \\
&= (\mathbf{A}^+)^T\mathbf{A}^T(\mathbf{B}^+)^T\mathbf{A}^T = (\mathbf{A}\mathbf{A}^+)^T(\mathbf{A}\mathbf{B}^+)^T = (\mathbf{A}\mathbf{A}^+)(\mathbf{A}\mathbf{B}^+) \\
&= (\mathbf{A}\mathbf{A}^+\mathbf{A})\mathbf{B}^+ = \mathbf{A}\mathbf{B}^+, \\
\mathbf{A}^+\mathbf{A} &= (\mathbf{A}^+\mathbf{A})^T = \mathbf{A}^T(\mathbf{A}^+)^T = (\mathbf{A}\mathbf{B}^+\mathbf{A})^T(\mathbf{A}^+)^T \\
&= \mathbf{A}^T(\mathbf{B}^+)^T\mathbf{A}^T(\mathbf{A}^+)^T = (\mathbf{B}^+\mathbf{A})^T(\mathbf{A}^+\mathbf{A})^T = (\mathbf{B}^+\mathbf{A})(\mathbf{A}^+\mathbf{A}) \\
&= \mathbf{B}^+(\mathbf{A}\mathbf{A}^+\mathbf{A}) = \mathbf{B}^+\mathbf{A}.
\end{aligned}$$

Now we have

$$\begin{aligned}\mathbf{A}^+ &= \mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+(\mathbf{A}\mathbf{A}^+) = \mathbf{A}^+(\mathbf{A}\mathbf{B}^+)\\ &= (\mathbf{A}^+\mathbf{A})\mathbf{B}^+ = (\mathbf{B}^+\mathbf{A})\mathbf{B}^+ = \mathbf{B}^+\mathbf{A}\mathbf{B}^+ = \mathbf{B}^+.\end{aligned}$$

$\square$

Existence of Pseudoinverse

Let

$$\mathbf{A} = \mathbf{U_A}\mathbf{\Sigma_A}\mathbf{V_A}^T$$

be a singular value decomposition of $\mathbf{A}$, where $\mathbf{U_A}^T\mathbf{U_A} = \mathbf{V_A}^T\mathbf{V_A} = \mathbf{I}_{r\times r}$. Then

$$\mathbf{A}^+ = \mathbf{V_A}\mathbf{\Sigma_A}^{-1}\mathbf{U_A}^T$$

is the pseudoinverse of A by checking

- $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{U_A}\mathbf{\Sigma_A}(\mathbf{V_A}^T\mathbf{V_A})\mathbf{\Sigma_A}^{-1}(\mathbf{U_A}^T\mathbf{U_A})\mathbf{\Sigma_A}\mathbf{V_A}^T = \mathbf{U_A}\mathbf{\Sigma_A}\mathbf{V_A}^T = \mathbf{A}$.
- $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{V_A}\mathbf{\Sigma_A}^{-1}(\mathbf{U_A}^T\mathbf{U_A})\mathbf{\Sigma_A}(\mathbf{V_A}^T\mathbf{V_A})\mathbf{\Sigma_A}^{-1}\mathbf{U_A}^T = \mathbf{V_A}\mathbf{\Sigma_A}^{-1}\mathbf{U_A}^T = \mathbf{A}^+$.
- Since $\mathbf{A}^+\mathbf{A} = \mathbf{V_A}\mathbf{\Sigma_A}^{-1}\mathbf{U_A}^T\mathbf{U_A}\mathbf{\Sigma_A}\mathbf{V_A}^T = \mathbf{V_A}\mathbf{V_A}^T$, $\mathbf{A}^+\mathbf{A}$ is symmetric.

- Since $\mathbf{A}\mathbf{A}^{+} = \mathbf{U}_{\mathbf{A}}\mathbf{\Sigma}_{\mathbf{A}}\mathbf{V}_{\mathbf{A}}^{T}\mathbf{V}_{\mathbf{A}}\mathbf{\Sigma}_{\mathbf{A}}^{-1}\mathbf{U}_{\mathbf{A}}^{T} = \mathbf{U}_{\mathbf{A}}\mathbf{U}_{\mathbf{A}}^{T}$, $\mathbf{A}\mathbf{A}^{+}$ is symmetric.

$\square$

The Pseudoinverse of an SPSD matrix

- $\mathbf{A}$: an $n \times n$ SPSD matrix.
- $\mathbf{A} = \mathbf{V_A \Lambda_A V_A^T}$: an SVD of $\mathbf{A}$, where $\mathbf{\Lambda_A}$ is an $r \times r$ diagonal matrix with all positive eigenvalues of $\mathbf{A}$ in the diagonal.
- $\mathbf{A}^+ = \mathbf{V_A \Lambda_A^{-1} V_A^T}$: the pseudoinverse of $\mathbf{A}$.

Other Types of Empirical Kernels

- K: a PDS kernel over an input space $\mathscr{I}$.
- $S = (\omega_1, \omega_2, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from $\mathscr{I}$ according to an unknown distribution P.
- $\mathbf{K} = [K(\omega_i, \omega_j)]$: the kernel matrix associated to the kernel K and the sample $S = (\omega_1, \omega_2, \ldots, \omega_m)$, which is SPSD.
  - $\mathbf{K} = \mathbf{V_K} \boldsymbol{\Lambda}_\mathbf{K} \mathbf{V}_\mathbf{K}^T$: an SVD of $\mathbf{K}$.
- $\mathbf{e}_i$: the ith standard unit vector in $\mathbb{R}^m$.
- $\Phi_S : \mathscr{I} \to \mathbb{R}^m$: the empirical kernel map associated to the kernel K and the sample S.
  - $\Phi_S(\omega_i) = \mathbf{K}\mathbf{e}_i$ for all $i \in [1, m]$.
- $\mathbf{K}^+ = \mathbf{V_K} \boldsymbol{\Lambda}_\mathbf{K}^{-1} \mathbf{V}_\mathbf{K}^T$: the pseudoinverse of $\mathbf{K}$.

- $\sqrt{\mathbf{K}^+} = \mathbf{V}_{\mathbf{K}} \sqrt{\mathbf{\Lambda}_{\mathbf{K}}^{-1}} \mathbf{V}_{\mathbf{K}}^T$: the square-root of the pseudoinverse $\mathbf{K}^+$ of $\mathbf{K}$.

- $\Psi_S(\omega) \triangleq \sqrt{\mathbf{K}^+} \Phi_S(\omega), \ \forall\, \omega \in \mathscr{I}$: a feature mapping which defines a type of empirical kernels by

$$\begin{aligned} K'_S(\omega, \omega') &= \Psi_S(\omega)^T \Psi_S(\omega') = \left(\sqrt{\mathbf{K}^+} \Phi_S(\omega)\right)^T \left(\sqrt{\mathbf{K}^+} \Phi_S(\omega')\right) \\ &= \Phi_S(\omega)^T \mathbf{K}^+ \Phi_S(\omega') \end{aligned}$$

  – The kernel matrix $\mathbf{K}'_S = [K'_S(\omega_i, \omega_j)]$ associated to the empirical kernel K'_S and the sample S is

$$\begin{aligned} K'_S(\omega_i, \omega_j) &= \Phi_S(\omega_i)^T \mathbf{K}^+ \Phi_S(\omega_j) = \mathbf{e}_i^T \mathbf{K} \mathbf{K}^+ \mathbf{K} \mathbf{e}_j = \mathbf{e}_i^T \mathbf{K} \mathbf{e}_j \\ &= K(\omega_i, \omega_j) \end{aligned}$$

  so that

$$\mathbf{K}'_S = \mathbf{K}.$$

- $\Omega_S(\omega) \triangleq \mathbf{K}^+\Phi_S(\omega),\ \forall\, \omega \in \mathscr{I}$: a feature mapping which defines a type of empirical kernels by

$$\begin{aligned} K''_S(\omega,\omega') &= \Omega_S(\omega)^T\Omega_S(\omega') = \left(\mathbf{K}^+\Phi_S(\omega)\right)^T\left(\mathbf{K}^+\Phi_S(\omega')\right) \\ &= \Phi_S(\omega)^T\mathbf{K}^+\mathbf{K}^+\Phi_S(\omega') \end{aligned}$$

  – The kernel matrix $\mathbf{K}''_S = [K''_S(\omega_i,\omega_j)]$ associated to the empirical kernel K''_S and the sample S is

$$\begin{aligned} K''_S(\omega_i,\omega_j) &= \Phi_S(\omega_i)^T\mathbf{K}^+\mathbf{K}^+\Phi_S(\omega_j) = \mathbf{e}_i^T\mathbf{K}\mathbf{K}^+\mathbf{K}^+\mathbf{K}\mathbf{e}_j \\ &= \mathbf{e}_i^T\mathbf{K}\mathbf{K}^+\mathbf{e}_j, \end{aligned}$$

  where $\mathbf{K}^+\mathbf{K}^+\mathbf{K} = \mathbf{V_K}\mathbf{\Lambda}_\mathbf{K}^{-1}\mathbf{V}_\mathbf{K}^T\mathbf{V_K}\mathbf{\Lambda}_\mathbf{K}^{-1}\mathbf{V}_\mathbf{K}^T\mathbf{V_K}\mathbf{\Lambda_K}\mathbf{V}_\mathbf{K}^T = \mathbf{V_K}\mathbf{\Lambda}_\mathbf{K}^{-1}\mathbf{V}_\mathbf{K}^T = \mathbf{K}^+$ so that

$$\mathbf{K}''_S = \mathbf{K}\mathbf{K}^+ = \mathbf{V_K}\mathbf{V}_\mathbf{K}^T,$$

  which is $\mathbf{I}_{m\times m}$ when $\mathbf{K}$ is invertible.

The Contents of This Lecture

- Positive definite symmetric (PDS) kernels
- Closure properties of PDS kernels
- Reproducing kernel Hilbert space (RKHS)
- SVMs with PDS kernels
- Conditionally negative definite symmetric (CNDS) kernels
- Sequence kernels

The Primal Problem for SVM with a PDS Kernel

- $\mathscr{I}$: the input space of all possible items, associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$, where P is unknown.
- $c : \mathscr{I} \to \{-1, +1\}$: a fixed but unknown concept.
- K: a PDS kernel over the input space $\mathscr{I}$.
- $\mathscr{F}$: a feature space, which is a Hilbert space over $\mathbb{R}$.
  - A commonly used feature space is the reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ associated to the PDS kernel K.
- Φ: a feature mapping from $\mathscr{I}$ to $\mathscr{F}$ such that for all ω, ω' in $\mathscr{I}$,

$$K(\omega, \omega') = \langle \Phi(\omega), \Phi(\omega') \rangle.$$

  - If $\mathscr{F}$ is the RKHS $\mathbb{H}$ associated to the PDS kernel K, we have

$$\forall\, \omega \in \mathscr{I}, \quad \Phi(\omega) = K(\omega, \cdot).$$

- $S = (\omega_1, \omega_2, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from the input space $\mathscr{I}$ according to the distribution P with labels $(c(\omega_1), c(\omega_2), \ldots, c(\omega_m))$.

The primal problem for SVM in a feature space $\mathscr{F}$ associated to the PDS kernel K is

$$\begin{array}{ll} \text{Minimize} & F(f, b, \eta) = \frac{1}{2}\|f\|^2_{\mathscr{F}} + C\sum_{i=1}^{m} \eta_i \\ \text{Subject to} & 1 - \eta_i - c(\omega_i)(\langle f, \Phi(\omega_i)\rangle + b) \leq 0, i = 1, \ldots, m \\ & -\eta_i \leq 0, i = 1, \ldots, m \\ & (f, b, \eta) \in \mathscr{F} \times \mathbb{R} \times \mathbb{R}^m. \end{array}$$

- How do we solve this primal problem when the feature space $\mathscr{F}$ is an infinite-dimensional Hilbert space ?

The Representer Theorem

Theorem 5.4: Let

- K: a PDS kernel over an input space $\mathscr{I}$.
- $\mathscr{F} = \mathbb{H}$: a feature space, which is the reproducing kernel Hilbert space (RKHS) associated to the PDS kernel K.
- $(\omega_1, \omega_2, \ldots, \omega_m)$: a given m-tuple over the input space $\mathscr{I}$.
- $G : \mathbb{R}^+ \rightarrow \mathbb{R}$: a non-decreasing function.
- $L : \mathbb{R}^m \rightarrow \mathbb{R} \cup \{\infty\}$: any function.

Any solution of the optimization problem

$$\text{Minimize}_{h\in\mathbb{H}}\ F(h) = G(\|h\|_{\mathbb{H}}) + L(h(\omega_1), h(\omega_2), \ldots, h(\omega_m))$$

admits a solution of the form

$$h^* = \sum_{i=1}^{m} \alpha_i K(\omega_i, \cdot),$$

for some real numbers $\alpha_i, i \in [1, m]$. If G is further assumed to be strictly increasing, then any solution has this form.

Proof.

- $\mathbb{H}_1 = \mathrm{Span}(\{K(\omega_i, \cdot), i \in [1, m]\})$: a finite-dimensional subspace of the RKHS $\mathbb{H}$, which is a closed subspace.
  - Closedness: if a sequence $\{h_n\}_{n=1}^{\infty}$ in $\mathbb{H}_1$ converges to an $h \in \mathbb{H}$, then h must be in $\mathbb{H}_1$.
- $\mathbb{H}_1^{\perp} = \{h \in \mathbb{H} : \langle h, h'\rangle = 0\ \forall\ h' \in \mathbb{H}_1\}$: the orthogonal complement of $\mathbb{H}_1$, which is a closed subspace of $\mathbb{H}$.

- Since $\mathbb{H}_1$ is closed, $\mathbb{H} = \mathbb{H}_1 \oplus \mathbb{H}_1^{\perp}$, i.e., $\mathbb{H}$ is the direct sum of $\mathbb{H}_1$ and $\mathbb{H}_1^{\perp}$, which means that for each $h \in \mathbb{H}$, there exist unique $h_1 \in \mathbb{H}_1$ and $h^{\perp} \in \mathbb{H}_1^{\perp}$ such that $h = h_1 + h^{\perp}$.
- Since G is non-decreasing, $G(\|h_1\|_{\mathbb{H}}) \leq G(\sqrt{\|h_1\|_{\mathbb{H}}^2 + \|h^{\perp}\|_{\mathbb{H}}^2}) = G(\|h\|_{\mathbb{H}})$.
- By the reproducing property, for all $i \in [1, m]$, $h(\omega_i) = \langle h, K(\omega_i, \cdot)\rangle = \langle h_1, K(\omega_i, \cdot)\rangle = h_1(\omega_i)$. Thus, $L(h(\omega_1), h(\omega_2), \ldots, h(\omega_m)) = L(h_1(\omega_1), h_1(\omega_2), \ldots, h_1(\omega_m))$.
- $F(h_1) \leq F(h)$ for all $h \in \mathbb{H}$, which proves the first part of the theorem.
- If G is further strictly increasing, then $F(h_1) < F(h)$ when $\|h^{\perp}\| > 0$ and any solution of the optimization problem must be in $\mathbb{H}_1$.

$\square$

Reformulation of Primal Problem for Kernel-Based SVM

- $\mathscr{I}$: the input space of all possible items, associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$, where P is unknown.
- $c : \mathscr{I} \to \{-1, +1\}$: a fixed but unknown concept.
- K: a PDS kernel over the input space $\mathscr{I}$.
- $\mathscr{F} = \mathbb{H}$: a feature space, which is the reproducing kernel Hilbert space (RKHS) $\mathbb{H}$ associated to the PDS kernel K with the feature mapping $\Phi : \mathscr{I} \to \mathbb{H}$ such that $\Phi(\omega) = K(\omega, \cdot)$.
- $S = (\omega_1, \omega_2, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from the input space $\mathscr{I}$ according to the distribution P with labels $(c(\omega_1), c(\omega_2), \ldots, c(\omega_m))$.

The primal problem for SVM in the RKHS feature space $\mathbb{H}$ associated to the PDS kernel K is

$$\begin{aligned} &\text{Minimize} && F(h,b,\eta) = \tfrac{1}{2}\|h\|_{\mathbb{H}}^2 + C\textstyle\sum_{i=1}^m \eta_i \\ &\text{Subject to} && 1-\eta_i - c(\omega_i)(\langle h, \Phi(\omega_i)\rangle + b) \le 0, i = 1,\dots,m \\ &&& -\eta_i \le 0, i=1,\dots,m \\ &&& (h,b,\eta) \in \mathbb{H}\times\mathbb{R}\times\mathbb{R}^m. \end{aligned}$$

which is equivalent to

$$\text{Minimize}_{h\in\mathbb{H}, b\in\mathbb{R}}\ \tilde{F}(h,b) = \frac{1}{2}\|h\|_{\mathbb{H}}^2 + C\sum_{i=1}^m \max(0, 1-c(\omega_i)(h(\omega_i)+b))$$

since

$$\eta_i \ge \max(0, 1-c(\omega_i)(h(\omega_i)+b)),\ \ i=1,2,\dots,m,$$

which is also equivalent to

$$\underset{b\in\mathbb{R}}{\text{Minimize}}\,\underset{h\in\mathbb{H}}{\text{Minimize}}\,\tilde{F}(h,b)=\frac{1}{2}\|h\|_{\mathbb{H}}^2+C\sum_{i=1}^{m}\max(0,1-c(\omega_i)(h(\omega_i)+b)).$$

By fixing $b\in\mathbb{R}$ and letting,

- $G(\|h\|_{\mathbb{H}})=\frac{1}{2}\|h\|_{\mathbb{H}}^2$ with $G(x)=\frac{1}{2}x^2$ strictly increasing;
- $L(h(\omega_1),h(\omega_2),\ldots,h(\omega_m))=$ $C\sum_{i=1}^{m}\max(0,1-c(\omega_i)(h(\omega_i)+b))$,

any solution of the optimization problem

$$\underset{h\in\mathbb{H}}{\text{Minimize}}\,\tilde{F}(h,b)=\frac{1}{2}\|h\|_{\mathbb{H}}^2+C\sum_{i=1}^{m}\max(0,1-c(\omega_i)(h(\omega_i)+b))$$

must be of the form $h^{*,b}=\sum_{i=1}^{m}\alpha_i^bK(\omega_i,\cdot)$ by the representer theorem.

Let

$$\begin{aligned}\mathbb{H}_S &\triangleq \mathrm{Span}\{K(\omega_j,\cdot), j=1,2,\ldots,m\}\\ &= \left\{\sum_{j=1}^{m} \alpha_j K(\omega_j,\cdot) \mid \alpha_j \in \mathbb{R},\ 1 \leq m \leq m\right\},\end{aligned}$$

which is a finite-dimensional Hilbert space. Then for each fixed $b \in \mathbb{R}$, we have

$$\begin{aligned}&\underset{h\in\mathbb{H}}{\mathrm{Minimize}}\ \tilde{F}(h,b) = \frac{1}{2}\|h\|_{\mathbb{H}}^2 + C\sum_{i=1}^{m}\max(0, 1-c(\omega_i)(h(\omega_i)+b))\\ \Leftrightarrow\quad &\underset{h\in\mathbb{H}_S}{\mathrm{Minimize}}\ \tilde{F}(h,b) = \frac{1}{2}\|h\|_{\mathbb{H}_S}^2 + C\sum_{i=1}^{m}\max(0, 1-c(\omega_i)(h(\omega_i)+b))\end{aligned}$$

and then

$$\underset{h\in\mathbb{H},b\in\mathbb{R}}{\text{Minimize}}\ \tilde{F}(h,b)=\frac{1}{2}\|h\|^2_{\mathbb{H}}+C\sum_{i=1}^{m}\max(0,1-c(\omega_i)(h(\omega_i)+b))$$

$$\Leftrightarrow\quad \underset{h\in\mathbb{H}_S,b\in\mathbb{R}}{\text{Minimize}}\ \tilde{F}(h,b)=\frac{1}{2}\|h\|^2_{\mathbb{H}_S}+C\sum_{i=1}^{m}\max(0,1-c(\omega_i)(h(\omega_i)+b)).$$

Thus the primal problem for SVM in the RKHS feature space $\mathbb{H}$ associated to the PDS kernel K is equivalent to

$$\begin{array}{ll}\text{Minimize} & F(h,b,\eta)=\frac{1}{2}\|h\|^2_{\mathbb{H}_S}+C\sum_{i=1}^{m}\eta_i\\ \text{Subject to} & 1-\eta_i-c(\omega_i)(\langle h,\Phi(\omega_i)\rangle+b)\le 0, i=1,\ldots,m\\ & -\eta_i\le 0, i=1,\ldots,m\\ & (h,b,\eta)\in\mathbb{H}_S\times\mathbb{R}\times\mathbb{R}^m.\end{array}$$

The Lagrangian Dual Problem for Kernel-Based SVM

$$
\begin{array}{ll}
\text{Maximize} & \theta(\lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j c(\omega_i) c(\omega_j) \langle \Phi(\omega_i), \Phi(\omega_j) \rangle \\
\text{Subject to} & \lambda_i \geq 0,\ C - \lambda_i \geq 0, i = 1, \ldots, m \\
 & \sum_{i=1}^{m} \lambda_i c(\omega_i) = 0 \\
 & \lambda \in \mathbb{R}^m
\end{array}
$$

or equivalently

$$
\begin{array}{ll}
\text{Maximize} & \theta(\lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j c(\omega_i) c(\omega_j) K(\omega_i, \omega_j) \\
\text{Subject to} & \lambda_i \geq 0,\ C - \lambda_i \geq 0, i = 1, \ldots, m \\
 & \sum_{i=1}^{m} \lambda_i c(\omega_i) = 0 \\
 & \lambda \in \mathbb{R}^m
\end{array}
$$

The Kernel-Based SVM Algorithm

- $S = (\omega_1, \omega_2, \ldots, \omega_m)$: a labeled training sample of size m with labels $(c(\omega_1), c(\omega_2), \ldots, c(\omega_m))$.
- h_S^{SVM}: the hypothesis returned by SVM,

$$
\begin{aligned}
h_S^{SVM}(\omega) &= \operatorname{sgn}\left(\sum_{i=1}^{m} \lambda_i^{SVM} c(\omega_i) \langle \Phi(\omega_i), \Phi(\omega) \rangle + b^{SVM} \right) \\
&= \operatorname{sgn}\left(\sum_{i=1}^{m} \lambda_i^{SVM} c(\omega_i) K(\omega_i, \omega) + b^{SVM} \right)
\end{aligned}
$$

- $b^{SVM} = c(\omega_j) - \sum_{i=1}^{m} \lambda_i^{SVM} c(\omega_i) \langle \Phi(\omega_i), \Phi(\omega_j) \rangle$ for any support vector $\Phi(\omega_j)$ with $0 < \lambda_j < C$.

Thus we have

$$
\begin{aligned}
& h_S^{SVM}(\omega) \\
= \quad & \operatorname{sgn}\left(c(\omega_j) + \sum_{i=1}^{m} \lambda_i^{SVM} c(\omega_i) \langle \Phi(\omega_i), \Phi(\omega) - \Phi(\omega_j) \rangle \right) \\
= \quad & \operatorname{sgn}\left(c(\omega_j) + \sum_{i=1}^{m} \lambda_i^{SVM} c(\omega_i) (K(\omega_i, \omega) - K(\omega_i, \omega_j)) \right)
\end{aligned}
$$

for any support vector $\Phi(\omega_j)$ with $0 < \lambda_j < C$.

The Kernel-Based SVM Soft Margin ρ_{SVM}

- $b^{SVM} = c(\omega_j) - c(\omega_j)\eta_j^{SVM} - \sum_{i=1}^m \lambda_i^{SVM} c(\omega_i)\langle \Phi(\omega_i), \Phi(\omega_j)\rangle$ for any support vector $\Phi(\omega_j)$, i.e., $\lambda_j^{SVM} > 0$. This implies

$$
\begin{aligned}
& \sum_{j=1}^m \lambda_j^{SVM} c(\omega_j) b^{SVM} \\
= \quad & \sum_{j=1}^m \lambda_j^{SVM} (1 - \eta_j^{SVM}) c(\omega_j)^2 \\
& - \sum_{j=1}^m \lambda_j^{SVM} c(\omega_j) \sum_{i=1}^m \lambda_i^{SVM} c(\omega_i) \langle \Phi(\omega_i), \Phi(\omega_j)\rangle.
\end{aligned}
$$

- Since $\sum_{j=1}^m \lambda_j^{SVM} c(\omega_j) = 0$ and

$\mathbf{w}^{SVM} = \sum_{i=1}^{m} \lambda_i^{SVM} c(\omega_i)\Phi(\omega_i)$, we have

$$\sum_{j=1}^{m} \lambda_j^{SVM}(1 - \eta_j^{SVM}) = \|\mathbf{w}^{SVM}\|^2.$$

- ${\rho_{SVM}}^2 = \frac{1}{\|\mathbf{w}^{SVM}\|^2} = \frac{1}{\sum_{j=1}^{m} \lambda_j^{SVM}(1-\eta_j^{SVM})}$.

Remarks

- Modulo the offset b, the hypothesis solution h_S^{SVM} of kernel-based SVMs can be written as a linear combination of the functions $K(\omega_i, \cdot)$, where ω_i is a sample point.
- This is in fact a general property that holds for a broad class of optimization problems by applying the representer theorem.

Stirling's Formula

For any positive integer n, we have [a]

$$\sqrt{2\pi n}\, n^{n+\frac{1}{2}} e^{-n} e^{\frac{1}{12n+1}} < n! < \sqrt{2\pi n}\, n^{n+\frac{1}{2}} e^{-n} e^{\frac{1}{12n}}.$$

Thus we have

$$\frac{2^{2n}}{\sqrt{\pi n}} e^{-\frac{1}{24n(24n+1)}} < \binom{2n}{n} = \frac{(2n)!}{n!n!} < \frac{2^{2n}}{\sqrt{\pi n}} e^{\frac{1}{24n(12n+1)}}.$$

---

[a] H. Robbins, "A Remark on Stirling's Formula," *The American Mathematical Monthly*, 62 (1), pp. 26-29, 1955.

Rademacher Complexity of Bounded-Kernel-Based Affine Hypotheses with Bounded Weight Vector and Bounded Offset

Theorem 5.5: Let

- $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$: a PDS kernel over the input space $\mathscr{I}$ such that $K(\omega, \omega) \leq r^2 \ \forall\, \omega \in \mathscr{I}$ for some $r > 0$.
- $\mathscr{F} = \mathbb{H}$: a feature space, which is the reproducing kernel Hilbert space (RKHS) associated to the PDS kernel K.
- $\Phi : \mathscr{I} \to \mathbb{H}$: a feature mapping such that $\Phi(\omega) = K(\omega, \cdot)$ for all $\omega \in \mathscr{I}$ with $\langle \Phi(\omega), \Phi(\omega') \rangle = K(\omega, \omega')$.
- $S = (\omega_1, \omega_2, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from the input space $\mathscr{I}$ according to an unknown distribution P.

- $\mathcal{H} = \{\omega \mapsto \langle f, \Phi(\omega)\rangle + b \mid f \in \mathbb{H} \text{ with } \|f\|_{\mathbb{H}} \leq \Lambda,\ |b| \leq r\Lambda\}$: the set of all affine functionals in the Hilbert space $\mathbb{H}$ with bounded weight vector and bounded offset for some $\Lambda > 0$.

Then the empirical Rademacher complexity of $\mathcal{H}$ w.r.t. the sample S can be upper bounded as follows:

$$\hat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{\Lambda\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + \frac{r\Lambda}{\sqrt{m}} \leq 2\sqrt{\frac{r^2\Lambda^2}{m}},$$

where $\mathbf{K}$ is the kernel matrix associated to the kernel K and the sample S and $\mathrm{tr}(\mathbf{K})$ is the trace of $\mathbf{K}$.

Proof.

$$
\begin{aligned}
\hat{\mathfrak{R}}_S(\mathcal{H}) &= \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(\omega_i) \\
&= \frac{1}{m2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{\|f\|_{\mathbb{H}} \le \Lambda, |b| \le r\Lambda} \sum_{i=1}^{m} \sigma_i (\langle f, \Phi(\omega_i) \rangle + b) \\
&= \frac{1}{m2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{\|f\|_{\mathbb{H}} \le \Lambda} \langle f, \sum_{i=1}^{m} \sigma_i \Phi(\omega_i) \rangle \\
&\quad + \frac{1}{m2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{|b| \le r\Lambda} b \sum_{i=1}^{m} \sigma_i .
\end{aligned}
$$

Now the first average is

$$
\begin{aligned}
& \frac{1}{m2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{\|f\|_{\mathbb{H}} \le \Lambda} \langle f, \sum_{i=1}^{m} \sigma_i \Phi(\omega_i) \rangle \\
\le \quad & \frac{\Lambda}{m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \frac{1}{2^m} \sqrt{\langle \sum_{i=1}^{m} \sigma_i \Phi(\omega_i), \sum_{i=1}^{m} \sigma_i \Phi(\omega_i) \rangle} \\
& \text{by Cauchy-Schwarz inequality and } \|f\|_{\mathbb{H}} \le \Lambda \\
\le \quad & \frac{\Lambda}{m} \sqrt{\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \frac{1}{2^m} \sum_{i,j=1}^{m} \sigma_i \sigma_j \langle \Phi(\omega_i), \Phi(\omega_j) \rangle} \\
& \text{since } f(x) = \sqrt{x} \text{ is a concave function on } [0, \infty) \\
\le \quad & \frac{\Lambda}{m} \sqrt{\sum_{i,j=1}^{m} K(\omega_i, \omega_j) \frac{1}{2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sigma_i \sigma_j}.
\end{aligned}
$$

Since

$$\frac{1}{2^m}\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}}\sigma_i\sigma_j=\begin{cases}0, & \text{if } i\neq j,\\ 1, & \text{if } i=j,\end{cases}$$

we have

$$\begin{aligned}
& \frac{1}{m2^m}\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}}\sup_{\|f\|_{\mathbb{H}}\leq\Lambda}\langle f,\sum_{i=1}^m\sigma_i\Phi(\omega_i)\rangle \\
\leq\quad & \frac{\Lambda}{m}\sqrt{\sum_{i=1}^m K(\omega_i,\omega_i)}=\frac{\Lambda\sqrt{\mathrm{tr}(\mathbf{K})}}{m} \\
\leq\quad & \frac{\Lambda}{m}\sqrt{mr^2}=\sqrt{\frac{\Lambda^2r^2}{m}}.
\end{aligned}$$

And the second average is

$$
\begin{aligned}
& \frac{1}{m2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \sup_{|b| \le r\Lambda} b \sum_{i=1}^{m} \sigma_i \\
= \quad & \frac{r\Lambda}{m2^m} \sum_{\sigma_1,\sigma_2,\ldots,\sigma_m \in \{-1,+1\}} \left| \sum_{i=1}^{m} \sigma_i \right| \\
= \quad & \frac{r\Lambda}{m2^m} \, 2 \sum_{i=0}^{\lfloor m/2 \rfloor} \binom{m}{i} (m - 2i).
\end{aligned}
$$

Since

$$
\begin{aligned}
2\sum_{i=0}^{\lfloor m/2\rfloor}\binom{m}{i}(m-2i) &= \begin{cases} 2n\binom{2n}{n}, & \text{if } m=2n, \\ 2(2n+1)\binom{2n}{n}, & \text{if } m=2n+1 \end{cases} \\
&\le \begin{cases} \frac{2n2^{2n}}{\sqrt{\pi n}} e^{\frac{1}{24n(12n+1)}}, & \text{if } m=2n, \\ \frac{2(2n+1)2^{2n}}{\sqrt{\pi n}} e^{\frac{1}{24n(12n+1)}}, & \text{if } m=2n+1, \end{cases} \\
&\le \frac{m2^m}{\sqrt{m}}
\end{aligned}
$$

by Stirlng's formula, we have

$$
\frac{1}{2^m}\sum_{\sigma_1,\sigma_2,\ldots,\sigma_m\in\{-1,+1\}}\sup_{|b|\le r\Lambda}\frac{1}{m}b\sum_{i=1}^{m}\sigma_i \quad\le\quad \frac{r\Lambda}{m2^m}\frac{m2^m}{\sqrt{m}} = \frac{r\Lambda}{\sqrt{m}}.
$$

Thus we have $\hat{\mathfrak{R}}_S(\mathcal{H}) \le \frac{\Lambda\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + \sqrt{\frac{r^2\Lambda^2}{m}} \le 2\sqrt{\frac{r^2\Lambda^2}{m}}$. $\square$

Remarks

- The trace of the kernel matrix $\mathbf{K}$ is an important quantity for controlling the empirical Rademacher complexity of bounded-kernel-based affine hypothesis sets.
- By averaging over all samples S, we have
$$\mathfrak{R}_m(\mathcal{H}) \le 2\sqrt{\frac{r^2\Lambda^2}{m}}.$$
- With the bounded kernel $K(\omega,\omega) \le r^2$ for all $\omega \in \mathscr{I}$ and a bounded weight vector $\|f\|_{\mathbb{H}} \le \Lambda$, we have
$$-r\Lambda \le \langle f, \Phi(\omega)\rangle \le r\Lambda$$
since $\|f\|_{\mathbb{H}} \le \Lambda$ and $\|\Phi(\omega)\|_{\mathbb{H}} = \sqrt{K(\omega,\omega)} \le \Lambda$ so that
$$b - r\Lambda \le h(\omega) = \langle f, \Phi(\omega)\rangle + b \le b + r\Lambda,\ \forall\, \omega \in \mathscr{I}.$$

- When either $b > r\Lambda$ or $b < -r\Lambda$, we have either $h(\omega) > 0$ for all $\omega \in \mathscr{I}$ or $h(\omega) < 0$ for all $\omega \in \mathscr{I}$. In either case, the affine classifier h becomes trivial.

- From the proof of Theorem 5.5, we have

$$\hat{\mathfrak{R}}_S(\mathcal{H}) \approx \frac{\Lambda}{m} \underset{\sigma}{E}[\| \sum_{i=1}^{m} \sigma_i \Phi(\omega_i) \|_{\mathbb{H}}] + \frac{r\Lambda}{\sqrt{(\pi/2)m}}$$

  and by the Khintchine-Kahane inequality in Theorem D.4, we have

$$\underset{\sigma}{E}[\| \sum_{i=1}^{m} \sigma_i \Phi(\omega_i) \|_{\mathbb{H}}] \geq \sqrt{\frac{1}{2} \underset{\sigma}{E}[\| \sum_{i=1}^{m} \sigma_i \Phi(\omega_i) \|_{\mathbb{H}}^2]} = \sqrt{\frac{\mathrm{tr}(\mathbf{K})}{2}}$$

  so that the empirical Rademacher complexity $\hat{\mathfrak{R}}_S(\mathcal{H})$ can also be lower bounded by $\frac{1}{\sqrt{2}} \frac{\Lambda\sqrt{\mathrm{tr}(\mathbf{K})}}{m} + \frac{r\Lambda}{\sqrt{(\pi/2)m}}$.

Margin-Based Generalization Bound for Bounded-Kernel-Based Affine Hypotheses with Bounded Weight Vector and Bounded Offset

Corollary 5.1: Let

- $\mathscr{I}$: the input space, associated with a probability space $(\mathscr{I}, \mathcal{F}, P)$.
- $c : \mathscr{I} \to \{-1, +1\}$: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $S = (\omega_1, \omega_2, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from the input space $\mathscr{I}$ according to the unknown distribution P with labels $(c(\omega_1), c(\omega_2), \ldots, c(\omega_m))$.
- $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$: a PDS kernel over the input space $\mathscr{I}$ such that $K(\omega, \omega) \leq r^2 \ \forall\, \omega \in \mathscr{I}$ for some $r > 0$.

- $\mathscr{F} = \mathbb{H}$: a feature space, which is the reproducing kernel Hilbert space (RKHS) associated to the PDS kernel K.
- $\Phi : \mathscr{I} \to \mathbb{H}$: a feature mapping such that $\Phi(\omega) = K(\omega, \cdot)$ for all $\omega \in \mathscr{I}$ with $\langle \Phi(\omega), \Phi(\omega') \rangle = K(\omega, \omega')$.
- $\mathcal{H} = \{\omega \mapsto \langle f, \Phi(\omega) \rangle + b \mid \|f\|_{\mathbb{H}} \leq \Lambda,\ |b| \leq r\Lambda\}$: the set of all affine functionals of the Hilbert space $\mathbb{H}$ with bounded weight vector and bounded offset.
  - It is clear that $\sup_{h \in \mathcal{H}} |h(\omega)| \leq 2r\Lambda < +\infty\ \forall\ \omega \in \mathscr{I}$.
- $\rho > 0$: a given margin.
- $L_\rho(y', y) = \Phi_\rho(y'y) : \mathbb{R} \times \mathbb{R} \to [0, 1]$: the ρ-margin loss function.
- $\hat{R}_{S,\rho}(h) = \frac{1}{m} \sum_{i=1}^{m} L_\rho(h(\omega_i), c(\omega_i)) = \frac{1}{m} \sum_{i=1}^{m} \Phi_\rho(h(\omega_i) c(\omega_i))$: the empirical ρ-margin loss of an affine hypothesis h in $\mathcal{H}$ w.r.t. the concept c on the sample S.

- $R(h) = \underset{\omega \sim P}{E}[1_{\mathrm{sgn}(h(\omega)) \neq c(\omega)}]$: the generalization error of an affine hypothesis $h \in \mathcal{H}$.

For any $\delta > 0$, with probability at least $1 - \delta$, all h in $\mathcal{H}$:

$$R(h) \leq \hat{R}_{S,\rho}(h) + 4\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}}$$

$$R(h) \leq \hat{R}_{S,\rho}(h) + 2\frac{\sqrt{\mathrm{tr}(\mathbf{K})\Lambda^2/\rho^2}}{m} + 2\sqrt{\frac{r^2\Lambda^2/\rho^2}{m}} + 3\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}.$$

Proof. This is a direct consequence of Theorems 5.5 and 4.4. □

The Contents of This Lecture

- Positive definite symmetric (PDS) kernels
- Closure properties of PDS kernels
- Reproducing kernel Hilbert space (RKHS)
- SVMs with PDS kernels
- Conditionally negative definite symmetric (CNDS) kernels
- Sequence kernels

Conditionally Negative Definite Symmetric (CNDS) Kernels

A kernel $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$ over an input space $\mathscr{I}$ is said to be conditionally negative-definite symmetric (CNDS) if

- it is symmetric, i.e., $K(\omega, \omega') = K(\omega', \omega)$ for all $\omega, \omega' \in \mathscr{I}$;
- for all m-tuple $(\omega_1, \omega_2, \ldots, \omega_m)$ over $\mathscr{I}$ and $\mathbf{c} \in \mathbb{R}^m$ with $\mathbf{1}^T\mathbf{c} = \sum_{i=1}^m c_i = 0$, the following holds:

  $$\mathbf{c}^T\mathbf{K}\mathbf{c} = \sum_{i,j=1}^{m} c_i K(\omega_i, \omega_j) c_j \leq 0,$$

  where $\mathbf{K} = [K(\omega_i, \omega_j)]$.

Remarks

- If a kernel K is PDS, then $-K$ is NDS and then CNDS. But the converse does not hold in general.
- In practice, a natural distance or metric is available for the learning task considered and can be used to define a similarity measure, i.e., a kernel.
- As an example, Gaussian kernels

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$

have the form $\exp(-d^2)$, where $d(\mathbf{x}, \mathbf{x}') = \frac{1}{\sqrt{2}\sigma}\|\mathbf{x} - \mathbf{x}'\|$ is a metric for the input vector space $\mathbb{R}^N$.

- Several natural questions arise such as:
  - What other PDS kernels can we construct from a metric d in a Hilbert space?
  - What technical condition should d satisfy to guarantee that $\exp(-d^2)$ is PDS?
- A natural mathematical definition that helps address these questions is that of conditional negative definite symmetric (CNDS) kernels.

Example 5.3: Squared Euclidean Distance - A CNDS Kernel

The squared Euclidean distance in an inner product space $\mathbb{H}_0$ over $\mathbb{R}$

$$K(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^2_{\mathbb{H}_0}$$

is a CNDS kernel over $\mathbb{H}_0$.

Proof. It is clear that K is symmetric. Let $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)$ be an m-tuple over $\mathbb{H}_0$ and $\mathbf{c} \in \mathbb{R}^m$ with $\mathbf{1}^T\mathbf{c} = 0$. Let $\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]$.

$$
\begin{aligned}
& \mathbf{c}^T\mathbf{K}\mathbf{c} \\
= \quad & \sum_{i,j=1}^{m} c_i K(\mathbf{x}_i, \mathbf{x}_j) c_j = \sum_{i,j=1}^{m} c_i \|\mathbf{x}_i - \mathbf{x}_j\|^2 c_j \\
= \quad & \sum_{i,j=1}^{m} c_i c_j (\|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2\langle \mathbf{x}_i, \mathbf{x}_j \rangle \\
= \quad & \sum_{j=1}^{m} c_j \sum_{i=1}^{m} c_i \|\mathbf{x}_i\|^2 + \sum_{i=1}^{m} c_i \sum_{j=1}^{m} c_j \|\mathbf{x}_j\|^2 - 2\langle \sum_{i=1}^{m} c_i \mathbf{x}_i, \sum_{j=1}^{m} c_j \mathbf{x}_j \rangle \\
\leq \quad & 0.
\end{aligned}
$$

$\square$

CNDS Kernels v.s. PDS Kernels

Theorem 5.6: Let K be a symmetric kernel over an input space $\mathscr{I}$. Given a fixed $\omega_0 \in \mathscr{I}$, define a kernel K' over $\mathscr{I}$ as follows:

$$K'(\omega,\omega') \triangleq K(\omega,\omega_0)+K(\omega_0,\omega')-K(\omega_0,\omega_0)-K(\omega,\omega')\ \forall\, \omega,\omega' \in \mathscr{I}.$$

Then, K is CNDS if and only if K' is PDS.

Proof. "$\Leftarrow$" Assume that K' is PDS. Let $(\omega_1, \omega_2, \ldots, \omega_m)$ be an m-tuple over $\mathscr{I}$ and $\mathbf{c} \in \mathbb{R}^m$ with $\mathbf{1}^T\mathbf{c} = \sum_{i=1}^m c_i = 0$. Then

$$\begin{aligned}
& \sum_{i,j=1}^{m} c_i K(\omega_i, \omega_j) c_j \\
= & \sum_{i,j=1}^{m} c_i c_j \left(K(\omega_i, \omega_0) + K(\omega_0, \omega_j) - K(\omega_0, \omega_0) - K'(\omega_i, \omega_j)\right) \\
= & \left(\sum_{j=1}^{m} c_j\right)\left(\sum_{i=1}^{m} c_i K(\omega_i, \omega_0)\right) + \left(\sum_{i=1}^{m} c_i\right)\left(\sum_{j=1}^{m} c_j K(\omega_0, \omega_j)\right) \\
& - \left(\sum_{i=1}^{m} c_i\right)^2 K(\omega_0, \omega_0) - \sum_{i,j=1}^{m} c_i K'(\omega_i, \omega_j) c_j \\
= & -\sum_{i,j=1}^{m} c_i K'(\omega_i, \omega_j) c_j \leq 0.
\end{aligned}$$

Thus K is CNDS. "$\Rightarrow$" Assume that K is CNDS. Let

$\alpha_1, \alpha_2, \ldots, \alpha_m$ be in $\mathbb{R}$. Let $\alpha_0 = -\sum_{i=1}^m \alpha_i$. Then we have

$$
\begin{aligned}
& \sum_{i,j=1}^{m} \alpha_i K'(\omega_i, \omega_j)\alpha_j \\
= \quad & \sum_{i,j=1}^{m} \alpha_i \alpha_j (K(\omega_i, \omega_0) + K(\omega_0, \omega_j) - K(\omega_0, \omega_0) - K(\omega_i, \omega_j)) \\
= \quad & \left(\sum_{j=1}^{m} \alpha_j\right) \left(\sum_{i=1}^{m} \alpha_i K(\omega_i, \omega_0)\right) + \left(\sum_{i=1}^{m} \alpha_i\right) \left(\sum_{j=1}^{m} \alpha_j K(\omega_0, \omega_j)\right) \\
& - \left(\sum_{i=1}^{m} \alpha_i\right)^2 K(\omega_0, \omega_0) - \sum_{i,j=1}^{m} \alpha_i K(\omega_i, \omega_j)\alpha_j \\
= \quad & -\sum_{i=1}^{m} \alpha_i \alpha_0 K(\omega_i, \omega_0) - \sum_{j=1}^{m} \alpha_0 \alpha_j K(\omega_0, \omega_j) - \alpha_0 \alpha_0 K(\omega_0, \omega_0) \\
& - \sum_{i,j=1}^{m} \alpha_i K(\omega_i, \omega_j)\alpha_j,
\end{aligned}
$$

which says that

$$\sum_{i,j=1}^{m} \alpha_i K'(\omega_i, \omega_j)\alpha_j = -\sum_{i,j=0}^{m} \alpha_i K(\omega_i, \omega_j)\alpha_j \geq 0$$

since $\sum_{i=0}^{m} \alpha_i = 0$. Thus K' is PDS. $\square$

CNDS Kernels v.s. Gaussian Kernels

Theorem 5.7: Let K be a symmetric kernel over an input space $\mathscr{I}$. Then K is CNDS if and only if $\exp(-tK)$ is PDS for any $t > 0$.

Proof. "$\Rightarrow$" First assume that K is CNDS. By Theorem 5.6,

$$K'(\omega, \omega') = K(\omega, \omega_0) + K(\omega_0, \omega') - K(\omega_0, \omega_0) - K(\omega, \omega'),\ \forall\, \omega, \omega' \in \mathscr{I},$$

is a PDS kernel for a fixed $\omega_0 \in \mathscr{I}$. Thus for any $t > 0$, we have

$$e^{-tK(\omega,\omega')} = e^{tK'(\omega,\omega')} \left(e^{-tK(\omega,\omega_0)} e^{-tK(\omega_0,\omega')} \right) e^{tK(\omega_0,\omega_0)}.$$

Since for any random sample $S = (\omega_1, \omega_2, \ldots, \omega_m)$ of size m and any real numbers $c_1, c_2, \ldots, c_m$, we have

$$\sum_{i,j=1}^{m} c_i c_j e^{-tK(\omega_i,\omega_0)} e^{-tK(\omega_0,\omega_j)} = \left(\sum_{i=1}^{m} c_i e^{-tK(\omega_i,\omega_0)} \right)^2 \geq 0$$

and then $e^{-tK(\omega,\omega_0)} e^{-tK(\omega_0,\omega')}$ is a PDS kernel. Also since

$e^{tK(\omega_0,\omega_0)}$ is a positive number and $e^{tK'(\omega,\omega')}$ is a PDS, e^{-tK} is PDS for any $t > 0$.

"$\Leftarrow$" Conversely, assume that e^{-tK} is PDS for any $t > 0$. Then $-e^{-tK}$ is NDS and then CNDS. It is easy to see that shifting by a constant and scaling by a positive constant $t > 0$ preserves the CNDS property so that $\frac{1-e^{-tK}}{t}$ is CNDS. Note that

$$\lim_{t\downarrow 0} \frac{e^{-tK(\omega,\omega')} - 1}{t - 0} = \frac{\partial e^{-tK(\omega,\omega')}}{\partial t}\Big|_{t=0} = -K(\omega,\omega'),\ \forall\, \omega,\omega' \in \mathscr{I}.$$

Now for any random sample $S = (\omega_1, \omega_2, \ldots, \omega_m)$ of size m and any real numbers $c_1, c_2, \ldots, c_m$ such that $\sum_{i=1}^m c_i = 0$, we have

$$\sum_{i,j=1}^m c_i \left(\frac{1 - e^{-tK(\omega_i,\omega_j)}}{t}\right) c_j \leq 0 \ \forall\ t > 0$$

so that

$$
\begin{aligned}
& \lim_{t \downarrow 0} \sum_{i,j=1}^{m} c_i \left(\frac{1 - e^{-tK(\omega_i, \omega_j)}}{t} \right) c_j \\
= & \sum_{i,j=1}^{m} c_i c_j \lim_{t \downarrow 0} \left(\frac{1 - e^{-tK(\omega_i, \omega_j)}}{t} \right) \\
= & \sum_{i,j=1}^{m} c_i c_j K(\omega_i, \omega_j) \leq 0,
\end{aligned}
$$

which shows that K is CNDS. $\square$

CNDS Kernels v.s. Metric

Theorem 5.8: Let $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$ be a CNDS kernel such that for all $\omega, \omega' \in \mathscr{I}$, $K(\omega, \omega') = 0$ iff $\omega = \omega'$. Then, there exist a Hilbert space $\mathbb{H}$ and a mapping $\Phi : \mathscr{I} \to \mathbb{H}$ such that for all $\omega, \omega' \in \mathscr{I}$,

$$K(\omega, \omega') = \|\Phi(\omega) - \Phi(\omega')\|_{\mathbb{H}}^2.$$

Thus, under the hypothesis of the theorem, $\sqrt{K}$ defines a metric in the input space $\mathscr{I}$.

Proof. Since K is a CNDS kernel, by Theorem 5.6,

$$K'(\omega, \omega') = \frac{1}{2}\left(K(\omega, \omega_0) + K(\omega_0, \omega') - K(\omega_0, \omega_0) - K(\omega, \omega')\right), \ \forall\, \omega, \omega' \in \mathscr{I},$$

is a PDS kernel for any $\omega_0 \in \mathscr{I}$. Let $\mathbb{H}$ be the RKHS of K' with a feature mapping $\Phi : \mathscr{I} \to \mathbb{H}$ such that $K'(\omega, \omega') = \langle \Phi(\omega), \Phi(\omega') \rangle$

for all $\omega, \omega' \in \mathscr{I}$. Since $K(\omega_0, \omega_0) = 0$, we have

$$\begin{aligned}
& \|\Phi(\omega) - \Phi(\omega')\|_{\mathbb{H}}^2 \\
= \quad & \langle \Phi(\omega) - \Phi(\omega'), \Phi(\omega) - \Phi(\omega') \rangle \\
= \quad & \langle \Phi(\omega), \Phi(\omega) \rangle + \langle \Phi(\omega'), \Phi(\omega') \rangle - 2\langle \Phi(\omega), \Phi(\omega') \rangle \\
= \quad & \frac{1}{2}(K(\omega, \omega_0) + K(\omega_0, \omega) - K(\omega, \omega) + K(\omega', \omega_0) + K(\omega_0, \omega') \\
& -K(\omega', \omega') - 2K(\omega, \omega_0) - 2K(\omega_0, \omega') + 2K(\omega, \omega')) \\
= \quad & K(\omega, \omega')
\end{aligned}$$

since $K(\omega, \omega) = K(\omega', \omega') = 0$. Now $\sqrt{K(\omega, \omega')} = \|\Phi(\omega) - \Phi(\omega')\| \geq 0$ and $\sqrt{K(\omega, \omega')} = 0$ iff $\omega = \omega'$. (This implies that Φ is one-to-one.) Since $\|\Phi(\omega) - \Phi(\omega')\| = \|\Phi(\omega') - \Phi(\omega)\|$ and $\|\Phi(\omega) - \Phi(\omega')\| \leq \|\Phi(\omega) - \Phi(\omega'')\| + \|\Phi(\omega'') - \Phi(\omega')\|$, $\sqrt{K}$ is a metric. $\square$

The Contents of This Lecture

- Positive definite symmetric (PDS) kernels
- Closure properties of PDS kernels
- Reproducing kernel Hilbert space (RKHS)
- SVMs with PDS kernels
- Conditionally negative definite symmetric (CNDS) kernels
- Sequence kernels

Motivations

- To construct PDS kernels, i.e., kinds of similarity measures, for sequences or strings of symbols.
- Applications to computational biology, natural language processing and document processing.
- Introduction to a general framework for sequence kernels, rational kernels.

Multisets

- Multiset (or bag) : a generalization of the concept of a set. Unlike a set where an element counts only one membership, an element of a multiset may count many, even infinitely many, memberships.
- For example, $\{a, a, b\}$, $\{a, b, b\}$ and $\{a, b\}$ are three different multisets although they are the same set.
- Like any set, the order of elements in listing a multiset does not matter. Thus $\{a, b, b\}$ and $\{b, a, b\}$ are the same multiset.
- The multiplicity of an element in a multiset is the count of memberships of the element in the multiset. For example, in the multiset $\{a, a, a, b, b\}$, the multiplicity of a is 3, while that of b is 2.

Definition 5.4: Weighted Transducers

A weighted transducer T is a 7-tuple $T = (\Sigma, \Delta, Q, I, F, E, \rho)$ where

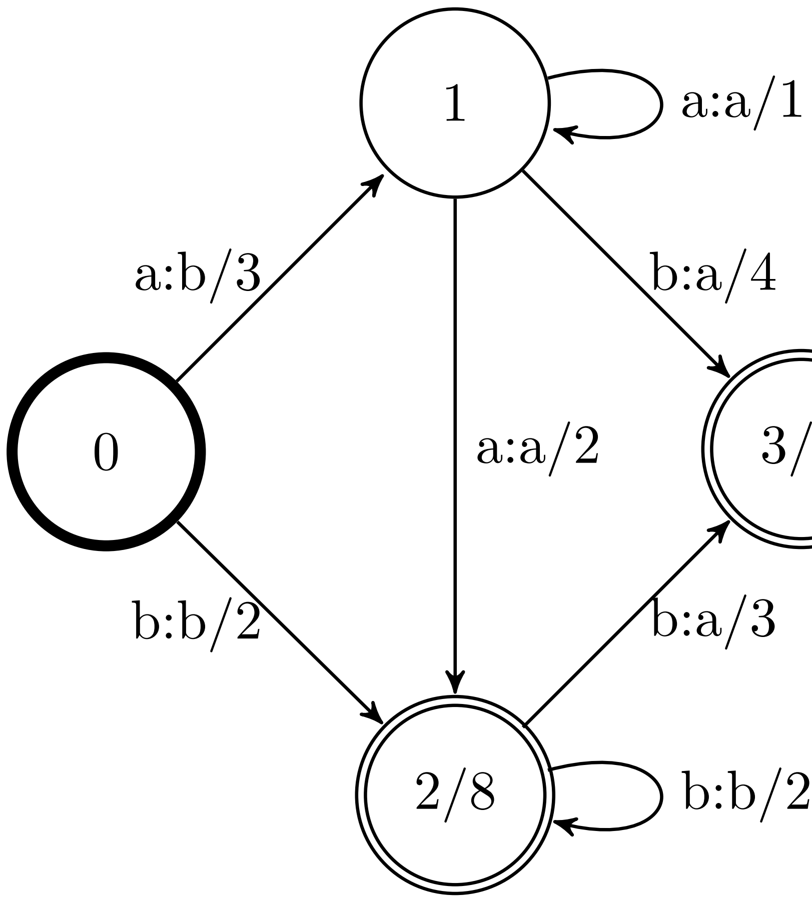
- Σ : a finite input alphabet,
  - An alphabet is a set of characters or a set of labels.
- Δ : a finite output alphabet,
- ϵ : the empty string or null label,
- Q : a finite set of states,
- $I \subseteq Q$: the set of initial states,
- $F \subseteq Q$: the set of final states,
- E : a finite multiset of transitions which are elements of $Q \times (\Sigma \cup \{\epsilon\}) \times (\Delta \cup \{\epsilon\}) \times \mathbb{R} \times Q$,
- $\rho : F \to \mathbb{R}$: a final weight function which maps F to $\mathbb{R}$.

State Transition Diagram of a Weighted Transducer

- Nodes with a bold circle : initial states,
- Nodes with double circles : final states,
  - The final weight $\rho(q)$ at a final state q is displayed after the slash.
- Node with a circle : intermediate states,
- Edges from a node to another node : transitions from a state to another state
  - Each edge is labeled by an input label and an output label separated by a colon delimiter, and a weight indicated after the slash separator.

Example : State Transition Diagram of a Weighted Transducer

Figure 5.3 of the *Foundations* textbook.



Terminologies for a Weighted Transducer $T = (\Sigma, \Delta, Q, I, F, E, \rho)$

- $E[q]$: the set of all outgoing edges from state q in a weighted transducer T,
- $i[e]$ and $o[e]$: the input label and output label of an edge e respectively,
- $p[e]$ and $n[e]$: the previous (origin) and next (destination) state of edge e respectively,
- $w[e]$: the weight of edge e.
- A path $\pi = e_1 e_2 \cdots e_k$: a sequence of finite number of edges with $n[e_i] = p[e_{i+1}]$ for $i \in [1, k-1]$

- $i[\pi]$: the input label of path π which is a string element of Σ^* obtained by concatenating input labels along the path π,

$$i[\pi] = i[e_1]i[e_2]\cdots i[e_k]$$

  - Σ^* : the collection of all strings of characters in the alphabet Σ, including the empty string ϵ.

- $o[\pi]$: the output label of path π which is a string element of Δ^* obtained by concatenating output labels along the path π,

$$o[\pi] = o[e_1]o[e_2]\cdots o[e_k]$$

- $p[\pi] \triangleq p[e_1]$ and $n[\pi] \triangleq n[e_k]$: the previous (origin) and next (destination) state of path π respectively,

- $w[\pi] = w[e_1]w[e_2]\cdots w[e_k](\rho(n[\pi])?)$: the weight of path π which is the product of the weights $w[e_i]$ of edges along the path and the final weight of the next state $n[\pi]$ if $n[\pi]$ is a final state.

The Weight of an Accepting Path

- An accepting path $\pi = e_1 e_2 \cdots e_k$: a path from an initial state to a final state
- The weight $w[\pi]$ of accepting path π : the result obtained by multiplying the weights of its constituent transitions and the weight of the final state of the path.

Weights of Input and Output String Pairs

- $T = (\Sigma, \Delta, Q, I, F, E, \rho)$: a weighted transducer;
- $x \in \Sigma^*$: an input string;
- $y \in \Delta^*$: an output string;
- $T(x, y)$: the sum of the weights of all accepting paths with input string x and output string y;
- $T : \Sigma^* \times \Delta^* \to \mathbb{R}$: assigning a weight to each pair $(x, y) \in \Sigma^* \times \Delta^*$ of input and output strings.
  - The mapping T can be represented as a real semi-infinite matrix $T = [T(x, y)]$ with Σ^* and Δ^* as row index set and column index set respectively.

- An example in Figure 5.3 : there are two accepting paths which generate the I-O string pair (aab, baa): $0 \to 1 \to 1 \to 3$ and $0 \to 1 \to 2 \to 3$ with weights $3 \cdot 1 \cdot 4 \cdot 2$ and $3 \cdot 2 \cdot 3 \cdot 2$ so that

$$T(aab, baa) = 3 \cdot 1 \cdot 4 \cdot 2 + 3 \cdot 2 \cdot 3 \cdot 2 = 60.$$

Composition of Weighted Transducers - As a Mapping

- $T_1 = (\Sigma, \Delta, Q_1, I_1, F_1, E_1, \rho_1)$: a weighted transducer
- $T_2 = (\Delta, \Omega, Q_2, I_2, F_2, E_2, \rho_2)$: a weighted transducer
- $T_1 \circ T_2 : \Sigma^* \times \Omega^* \to \mathbb{R}$: the composition of two mappings $T_1 : \Sigma^* \times \Delta^* \to \mathbb{R}$ and $T_2 : \Delta^* \times \Omega^* \to \mathbb{R}$ defined by

$$(T_1 \circ T_2)(x, y) \triangleq \sum_{z \in \Delta^*} T_1(x, z)\ T_2(z, y)\ \forall\ x \in \Sigma^*,\ y \in \Omega^*.$$

  – With matrix representation, the mapping $T_1 \circ T_2$ corresponds to a real semi-infinite matrix which is just the matrix multiplication of the two real semi-infinite matrices corresponding to the two mappings T_1 and T_2,

$$[T_1 \circ T_2(x, y)] = [T_1(x, z)][T_2(z, y)].$$

Computation of $(T_1 \circ T_2)(x, y)$

- $T_1 = (\Sigma, \Delta, Q_1, I_1, F_1, E_1, \rho_1)$: a weighted transducer
- $T_2 = (\Delta, \Omega, Q_2, I_2, F_2, E_2, \rho_2)$: a weighted transducer
- Assumption : each edge in T_1 or in T_2 is ϵ-free, i.e., the null label ϵ does not appear as the input label of an edge of T_1 or T_2 nor as the output label of an edge of T_1 or T_2
- $x \in \Sigma^*, z \in \Delta^*, y \in \Omega^*$: strings of length n
- (x, z) : an I-O string pair generated by k accepting paths in the weighted transducer T_1, $\pi^{(i)} = e_1^{(i)} e_2^{(i)} \cdots e_n^{(i)}$, $i \in [1, k]$
- (z, y) : an I-O string pair generated by m accepting paths in the weighted transducer T_2, ${\pi'}^{(j)} = {e'_1}^{(j)} {e'_2}^{(j)} \cdots {e'_n}^{(j)}$, $j \in [1, m]$

Now we have

$$
\begin{aligned}
& T(x,z)T(z,y) \\
= \quad & \sum_{i=1}^{k} w[\pi^{(i)}] \sum_{j=1}^{m} w[{\pi'}^{(j)}] = \sum_{i=1}^{k}\sum_{j=1}^{m} w[\pi^{(i)}]w[{\pi'}^{(j)}] \\
= \quad & \sum_{i=1}^{k}\sum_{j=1}^{m}(w[e_1^{(i)}]w[{e'_1}^{(j)}])\cdots(w[e_n^{(i)}]w[{e'_n}^{(j)}])(\rho_1(n[e_n^{(i)}])\rho_2(n[{e'_n}^{(j)}])).
\end{aligned}
$$

It is clear that for each $l \in [1,n]$, we have $o[e_l^{(i)}] = i[{e'_l}^{(j)}]$ which suggests to define the concatenation $e \wedge e'$ of an edge e in T_1 and an edge e' in T_2 whenever $o(e) = i(e')$ to be an edge in $(Q_1 \times Q_2) \times (\Sigma \cup \{\epsilon\}) \times (\Omega \cup \{\epsilon\}) \times \mathbb{R} \times (Q_1 \times Q_2)$ such that

- $p[e \wedge e'] = (p[e], p[e'])$, $n[e \wedge e'] = (n[e], n[e'])$,
- $i[e \wedge e'] = i[e]$, $o[e \wedge e'] = o[e']$,
- $w[e \wedge e'] = w[e]w[e']$.

Now we have

$$\begin{aligned} & T(x,z)T(z,y) \\ = & \sum_{i=1}^{k}\sum_{j=1}^{m} w[e_1^{(i)} \wedge {e'_1}^{(j)}] \cdots w[e_n^{(i)} \wedge {e'_n}^{(j)}]\rho(n[e_n^{(i)} \wedge {e'_n}^{(j)}]). \end{aligned}$$

It can be seen that for each $i \in [1,k]$ and each $j \in [1,m]$, $(e_1^{(i)} \wedge {e'_1}^{(j)})(e_2^{(i)} \wedge {e'_2}^{(j)}) \cdots (e_n^{(i)} \wedge {e'_n}^{(j)})$ is a path with "initial state" $p[e_1^{(i)} \wedge {e'_1}^{(j)}] = (p[e_1^{(i)}], p[{e'_1}^{(j)}]) \in I_1 \times I_2$ and finial state $n[e_n^{(i)} \wedge {e'_n}^{(j)}] = (n[e_n^{(i)}], n[{e'_n}^{(j)}]) \in F_1 \times F_2$ with final weight

$$\rho(n[e_n^{(i)} \wedge {e'_n}^{(j)}]) \triangleq \rho_1(n[e_n^{(i)}])\rho_2(n[{e'_n}^{(j)}])$$

since for all $l \in [1, n-1]$,

$$n[e_l^{(i)} \wedge {e'_l}^{(j)}] = (n[e_l^{(i)}], n[{e'_l}^{(j)}]) = (p[e_{l+1}^{(i)}], p[{e'_{l+1}}^{(j)}]) = p[e_{l+1}^{(i)} \wedge {e'_{l+1}}^{(j)}].$$

- The discussion in above suggests to define a weighted transducer as the composition of T_1 and T_2.

Composition of Weighted Transducers - As a Transducer

- $T_1 = (\Sigma, \Delta, Q_1, I_1, F_1, E_1, \rho_1)$: a weighted transducer
- $T_2 = (\Delta, \Omega, Q_2, I_2, F_2, E_2, \rho_2)$: a weighted transducer
- Assumption : the null label ϵ does not appear as the input label of an edge of T_1 nor as the output label of an edge of T_2
- $T_1 \circ T_2 = (\Sigma, \Omega, Q, I, F, E, \rho))$: the composition of two transducers T_1 and T_2 as a weighted transducer with
  - $Q \subseteq Q_1 \times Q_2$;
  - $I = I_1 \times I_2 \subseteq Q$;
  - $F = Q \cap (F_1 \times F_2)$;
  - $E = \biguplus_{\substack{(q_1,a,b,w_1,q_2)\in E_1 \\ (q_1',b,c,w_2,q_2')\in E_2}} \{((q_1, q_1'), a, c, w_1 w_2, (q_2, q_2'))\}$,
    * $\biguplus$: the standard join operation of multisets as in

$\{1,2\} \uplus \{1,3,3\} = \{1,1,2,3,3\}$, and preserves the multiplicity of transitions.

– $\rho : F \to \mathbb{R}$ with the final weight $\rho(q)$ at a final state $q = (q_1, q_2)$ to be $\rho(q) = \rho_1(q_1)\rho_2(q_2)$.

An Algorithm for Weighted Composition $T_1 \circ T_2$

1. $Q \leftarrow I_1 \times I_2, \;\; I \leftarrow \emptyset, \;\; F \leftarrow \emptyset, \quad E \leftarrow \emptyset$
2. $\mathcal{Q} \leftarrow I_1 \times I_2$ % a queue containing the set of pairs of states
% yet to be examined
3. **while** $\mathcal{Q} \neq \emptyset$ **do**
4. $\quad q = (q_1, q_2) \leftarrow \mathrm{Head}(\mathcal{Q})$
5. $\quad$ Dequeue($\mathcal{Q}$)
6. $\quad$ **if** $q \in I_1 \times I_2$ **then**
7. $\qquad I \leftarrow I \cup \{q\}$
8. $\quad$ **if** $q \in F_1 \times F_2$ **then**
9. $\qquad F \leftarrow F \cup \{q\}$
10. $\qquad \rho(q) \leftarrow \rho_1(q_1) \cdot \rho_2(q_2)$

for each $(e_1, e_2) \in E[q_1] \times E[q_2]$ such that $o[e_1] = i[e_2]$ **do**

  if $q' = (n[e_1], n[e_2]) \notin Q$ **then**

    $Q \leftarrow Q \cup \{q'\}$

    Enqueue$(\mathcal{Q}, q')$

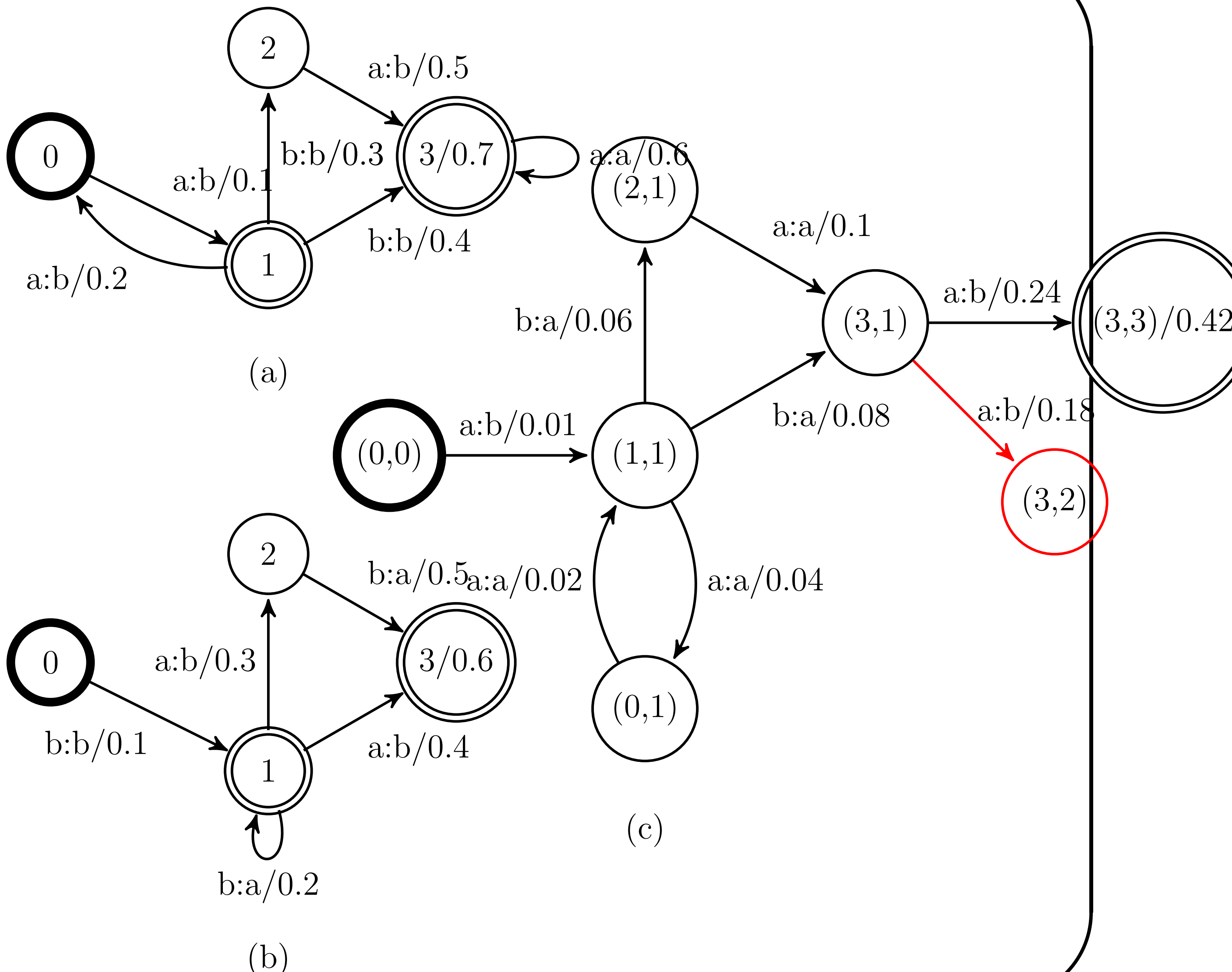
  $E \leftarrow E \biguplus \{(q, i[e_1], o[e_2], w[e_1] \cdot w[e_2], q')\}$

return T

where we have

- $E[q_i]$: sets of all edges emitting from state q_i in T_i,
- $i[e]$ and $o[e]$: the input label and output label of an edge e respectively,
- $p[e]$ and $n[e]$: the previous (origin) and next (destination) state of edge e respectively,
- $w[e]$: the weight of edge e.

Figure 5.4: Composition of Two Weighted Transducers


0
1
2
3/0.7
a:b/0.1
a:b/0.2
b:b/0.3
b:b/0.4
a:b/0.5
a:a/0.6
(a)
0
1
2
3/0.6
b:b/0.1
b:a/0.2
a:b/0.3
a:b/0.4
b:a/0.5
(b)
(0,0)
(1,1)
(0,1)
(2,1)
(3,1)
(3,2)
(3,3)/0.42
a:b/0.01
a:a/0.02
a:a/0.04
b:a/0.06
b:a/0.08
a:a/0.1
a:b/0.24
a:b/0.18
(c)

Remarks

- Special care should be taken when T_1 or T_2 is not ϵ-free since when T_1 admits output ϵ labels or T_2 input ϵ labels, the algorithm described in above may create redundant ϵ-paths, which would lead to an incorrect result.
- The weight of the matching paths of the original transducers would be counted p times, where p is the number of redundant paths in the result of composition.
- To avoid with this problem, all but one ϵ-path must be filtered out of the composite transducer.
- Remarkably, that filtering mechanism itself can be encoded as a finite-state transducer F.

Filtering of Redundant ϵ-Paths in Composition

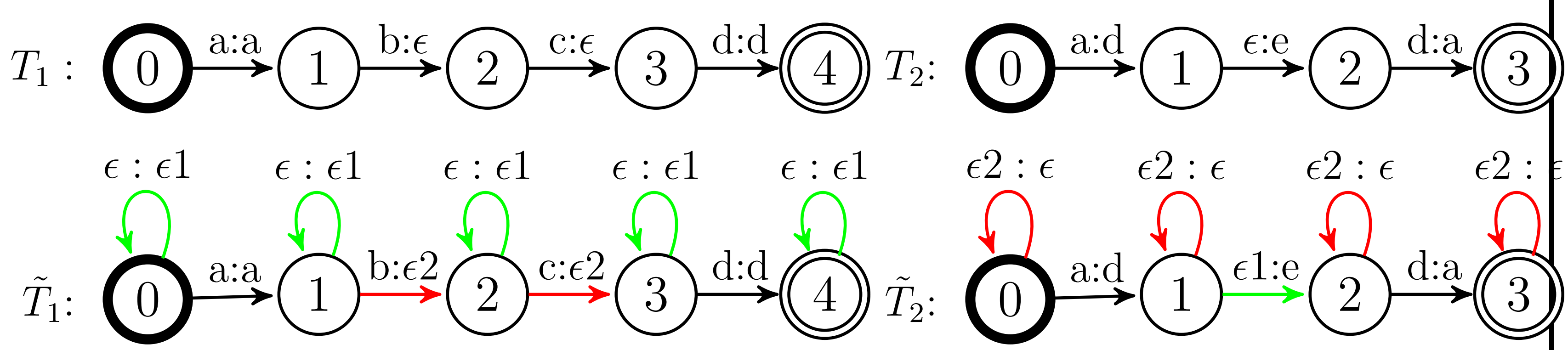
1. Augment T_1 and T_2 with auxiliary symbols that make the semantics of ϵ explicit.
2. $\tilde{T}_1$ and $\tilde{T}_2$: the weighted transducers obtained from T_1 and T_2 respectively by replacing the output (respectively input) ϵ labels with ϵ_2 (respectively ϵ_1) as illustrated by Figure 5.5.
3. Matching with the symbol ϵ_1 corresponds to remaining at the same state of T_1 and taking a transition of T_2 with input ϵ.
4. Matching with the symbol ϵ_2 corresponds to remaining at the same state of T_2 and taking a transition of T_1 with output ϵ.
5. The filter transducer F disallows a matching (ϵ_2, ϵ_2) immediately after (ϵ_1, ϵ_1) since this can be done instead via (ϵ_2, ϵ_1).

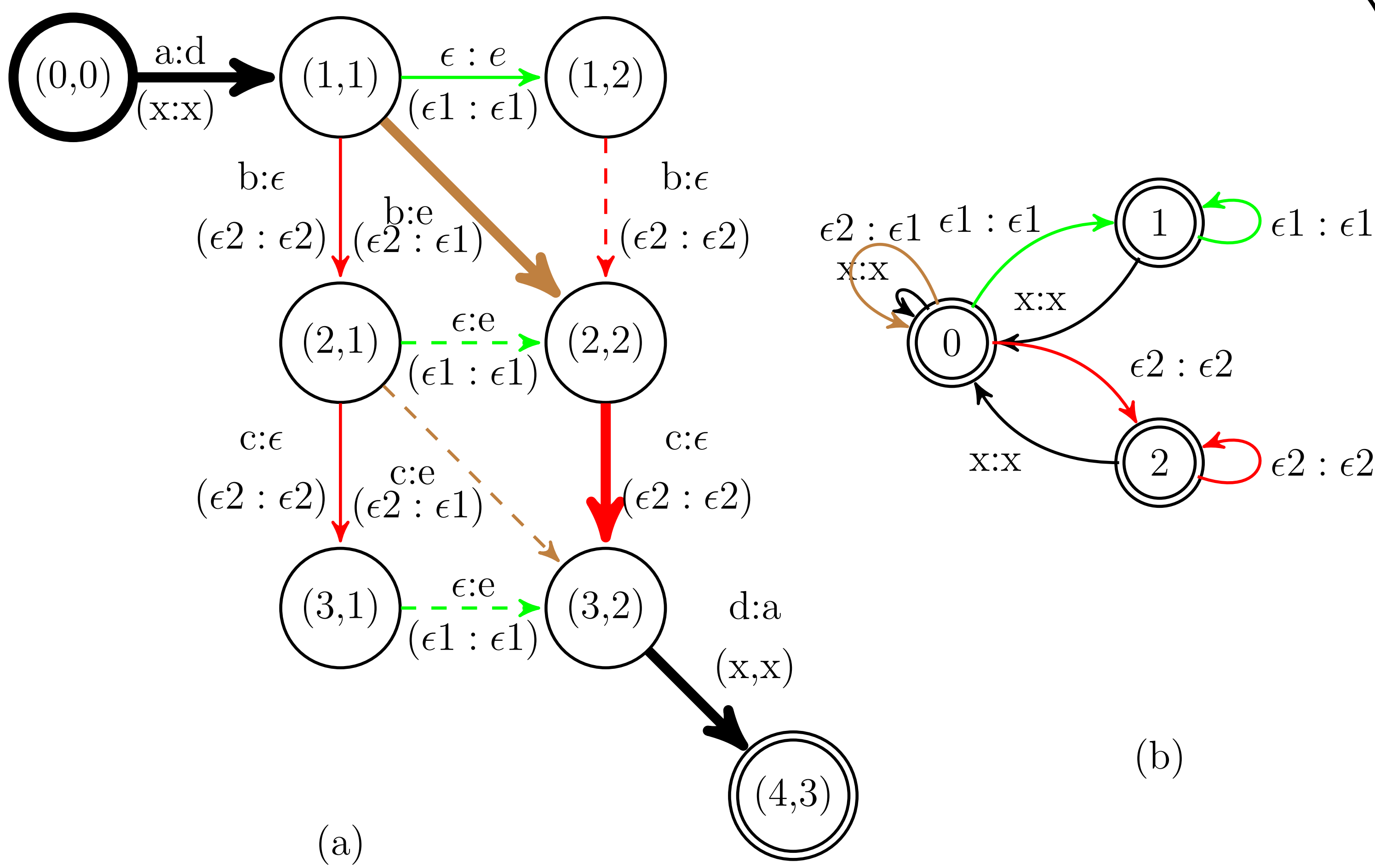
6. F also disallows a matching (ϵ_1, ϵ_1) immediately after (ϵ_2, ϵ_2).
7. Similarly, a matching (ϵ_1, ϵ_1) immediately followed by (ϵ_2, ϵ_1) is not permitted by the filter F since a path via the matchings $(\epsilon_2, \epsilon_1)(\epsilon_1, \epsilon_1)$ is possible.
8. And $(\epsilon_2, \epsilon_2)(\epsilon_2, \epsilon_1)$ is also ruled out.
9. Thus the filter transducer F is precisely a finite automaton over pairs accepting the complement of the language

   $$L = \sigma^*(\epsilon_1, \epsilon_1)(\epsilon_2, \epsilon_2) + (\epsilon_2, \epsilon_2)(\epsilon_1, \epsilon_1) + (\epsilon_1, \epsilon_1)(\epsilon_2, \epsilon_1) + (\epsilon_2, \epsilon_2)(\epsilon_2, \epsilon_1)\sigma^*$$

   where $\sigma = \{(\epsilon_1, \epsilon_1), (\epsilon_2, \epsilon_2), (\epsilon_2, \epsilon_1), x\}$.
10. Thus, the filter F guarantees that exactly one ϵ-path is allowed in the composition of each ϵ-sequence.
11. It is now legitimate to use the ϵ-free composition algorithm described in above to compute $\tilde{T}_1 \circ F \circ \tilde{T}_2$.

Figure 5.5: Dealing with Redundant ϵ-paths in Composition





(a) A straightforward generalization of the ϵ-free case would generate all the paths from $(1,1)$ to $(3,2)$ when composing T_1 and T_2 and may produce an incorrect result.

(b) Filter transducer F, where the shorthand x is used to represent an element of Σ.

Definition 5.5: Rational Kernels

A kernel $K : \Sigma^* \times \Sigma^* \to \mathbb{R}$ is said to be rational if it coincides with the mapping defined by some weighted transducer U: for all $x, y \in \Sigma^*$,

$$K(x, y) = U(x, y).$$

- Assumption : the transducer U does not admit any ϵ-cycle with non-zero weight, otherwise the kernel value is infinite for some pairs.
  - A cycle π is a path with $p[\pi] = n[\pi]$. An ϵ-cycle is a cycle with both input and output label equal to ϵ.
- For rational kernels, there exists a general and efficient computation algorithm.

Computation of $U(x, y)$

- x : a string in Σ^*;
- T_x : a weighted transducer with just one accepting path whose input and output labels are both x and its weight equal to one.
  - T_x can be straightforwardly constructed from x in linear time $O(|x|)$.

Step 1: Compute $V = T_x \circ U \circ T_y$ using the composition algorithm in time $O(|U||T_x||T_y|)$.

Step 2: Compute the sum of the weights of all accepting paths of V using a general shortest-distance algorithm in time $O(|V|)$.

- Since U admits no ϵ-cycle, V is acyclic, and this step can be performed in linear time.

The Inverse of a Weighted Transducer

For any weighted transducer T, let T^{-1} denote the inverse of T, that is the transducer obtained from T by swapping the input and output labels of every transition. For all $x, y \in \Sigma^*$, we have

$$T^{-1}(x, y) = T(y, x).$$

A Construction of PDS Rational Kernels

Theorem 5.3: For any weighted transducer $T = (\Sigma, \Delta, Q, I, F, E, \rho)$, the composite mapping $K = T \circ T^{-1}$ is a PDS rational kernel over Σ^*.

Proof.

- By definition, for all $x, y \in \Sigma^*$, we have

$$K(x, y) = \sum_{z \in \Delta^*} T(x, z) T^{-1}(z, y) = \sum_{z \in \Delta^*} T(x, z) T(y, z).$$

- K is the pointwise limit of the kernel sequence $\{K_n\}_{n=1}^{\infty}$ defined by: for all $n \in \mathbb{N}$ and $x, y \in \Sigma^*$,

$$K_n(x, y) \triangleq \sum_{|z| \leq n} T(x, z) T(y, z),$$

where the sum runs over all sequences in Σ^* of length $\leq n$.

- K_n is PDS since its corresponding kernel matrix $\mathbf{K}_n$ for any sample $S = (x_1, ..., x_m)$ drawn from Σ^* is SPSD since

$$\mathbf{K}_n = AA^T$$

with

$$A = [K_n(x_i, z_j)], \;\; i \in [1, m] \;\; \text{and} \;\; j \in [1, N],$$

where $z_1, \ldots, z_N$ is some arbitrary enumeration of the set of strings in Σ^* with length at most n.

- Thus, K is PDS as the pointwise limit of the sequence of PDS kernels $\{K_n\}_{n \in \mathbb{N}}$. $\square$

Bigram Transducers

- Σ : a finite alphabet of items
  - Items may be characters, letters, phonemes, syllables, words, DNA bases or amino acids.
- $z = z_1 z_2 \in \Sigma \times \Sigma$: a bigram
- T_{bigram} : the bigram transducer over Σ such that for each string $x \in \Sigma^*$ and each bigram $z = z_1 z_2$,

  $T_{\text{bigram}}(x, z) =$ the number of occurrences of the bigram z in x

Gappy-Bigram Transducers

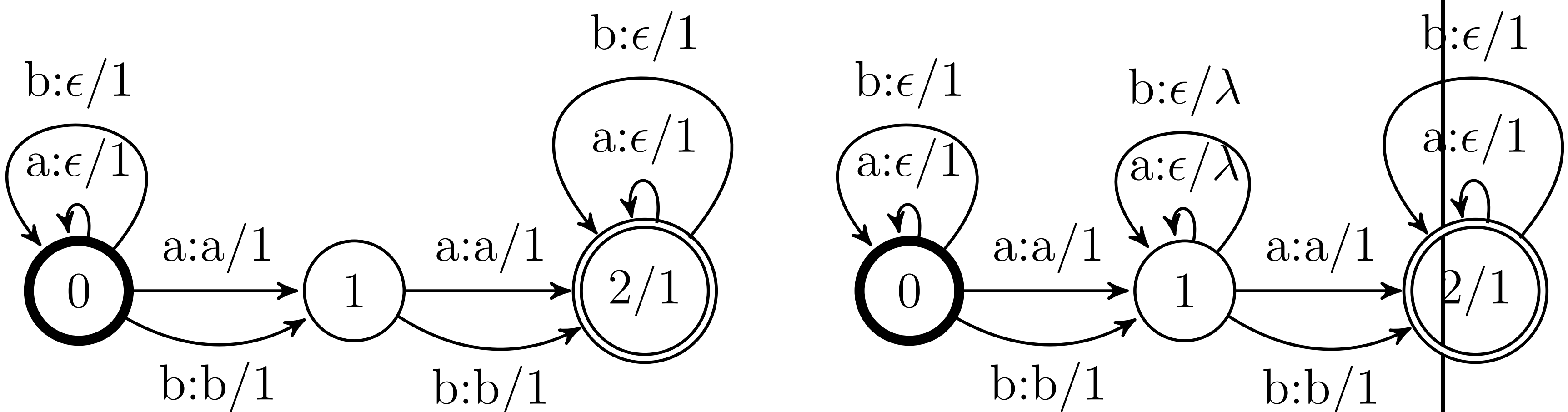
- Σ : a finite alphabet of items
- $z_1 u z_2 \in \Sigma \times \Sigma^* \times \Sigma$: a gappy bigram with gap u and gap penalty $\lambda^{|u|}$, where $\lambda \in (0, 1)$
- $T_{\text{gappy\_bigram}}$: the gappy_bigram transducer over Σ such that for each string $x \in \Sigma^*$ and each bigram $z = z_1 z_2$,

$$\begin{aligned} & T_{\text{gappy\_bigram}}(x, z) \\ = \; & \text{the sum of the number of occurrences of the gappy\_bigrams} \\ & z_1 u z_2 \text{ in } x \text{ weighted by the gap penalty } \lambda^{|u|} \text{ over all } u \in \Sigma^* \end{aligned}$$

Figure 5.6: Bigram and Gappy_Bigram Transducers

- $\Sigma = \{a, b\}$.

Left: Bigram transducer; Right: Gappy_bigram transducer



Example 5.5: Bigram and Gappy_Bigram Sequence Kernels

- Σ : a finite alphabet
- $K_{\text{bigram}} = T_{\text{bigram}} \circ T_{\text{bigram}}^{-1}$: the bigram kernel over Σ such that for any two strings x, y in Σ^*,

$$
\begin{aligned}
& K_{\text{bigram}}(x, y) \\
= \; & \sum_{z \in \Sigma^2} T_{\text{bigram}}(x, z) T_{\text{bigram}}(y, z) \\
= \; & \text{the sum of the product of the counts of} \\
& \text{all bigrams in } x \text{ and } y
\end{aligned}
$$

- $K_{\text{gappy\_bigram}} = T_{\text{gappy\_bigram}} \circ T^{-1}_{\text{gappy\_bigram}}$: the gappy_bigram kernel over Σ such that for any two strings x, y in Σ^*,

$$
\begin{aligned}
& K_{\text{gappy\_bigram}}(x, y) \\
= \quad & \sum_{z \in \Sigma^2} T_{\text{gappy\_bigram}}(x, z) T_{\text{gappy\_bigram}}(y, z) \\
= \quad & \text{the sum of the product of the gap-penalized counts of} \\
& \text{all bigrams in } x \text{ and } y
\end{aligned}
$$

Remarks

- Can we generalize the construction of bigram and gappy_bigram transducers to count the number of occurrences of certain patterns over an alphabet Σ and use them to define a PDS rational kernel ?
- The collection of those patterns is said to be a (formal) language over the alphabet Σ.
- Very often, it is a finite collection of patterns so that it is a regular language.
- Every regular language can be accepted by a finite automaton.

Regular Languages

The collection of regular languages over an alphabet Σ is defined recursively as follows:

- The empty language $\emptyset$ and the empty string language $\{\epsilon\}$ are regular languages.
- For each $a \in \Sigma$, the singleton language $\{a\}$ is a regular language.
- If A and B are regular languages, then $A \cup B$ (union), $A \bullet B$ (concatenation), and A^* (Kleene star) are regular languages.
  - $A \bullet B = \{ab \mid a \in A \text{ and } b \in B\}$.
  - $A^* = \{\epsilon\} \cup \{a_1 a_2 \cdots a_n \mid a_i \in A \; \forall\, i \in [1, n] \; \forall\, n \geq 1\}$.
- No other languages over Σ are regular.

Finite Automata and Regular Languages

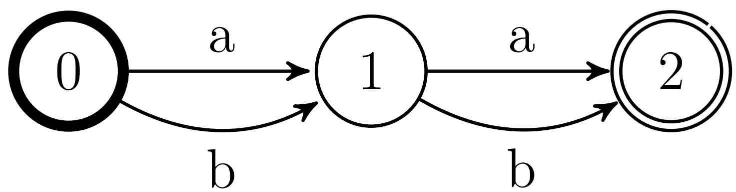
- A finite automaton A is a 5-tuple $A = (\Sigma, Q, I, F, E)$, where
  - Σ : a finite alphabet,
  - Q : a finite set of states,
  - $I \subseteq Q$: the set of initial states,
  - $F \subseteq Q$: the set of final states,
  - E : a finite set of transitions which are elements of $Q \times (\Sigma \cup \{\epsilon\}) \times Q$
- An accepting path : a path from an initial state to a final state in A.
- An accepted string : a string in Σ^* which labels an accepting path in A.

- $L(A) \subseteq \Sigma^*$: the set of all accepted strings by A.
  - $L(A)$ is called the language accepted by A and must be a regular language.

State Transition Diagram of an Automaton

- Nodes with a bold circle : initial states,
- Nodes with double circles : final states,
- Node with a circle : intermediate states,
- Edges from a node to another node : transitions from a state to another state
  - Each Edge is labeled by a label in $\Sigma \cup \{\epsilon\}$.

Example : A Finite Automaton X

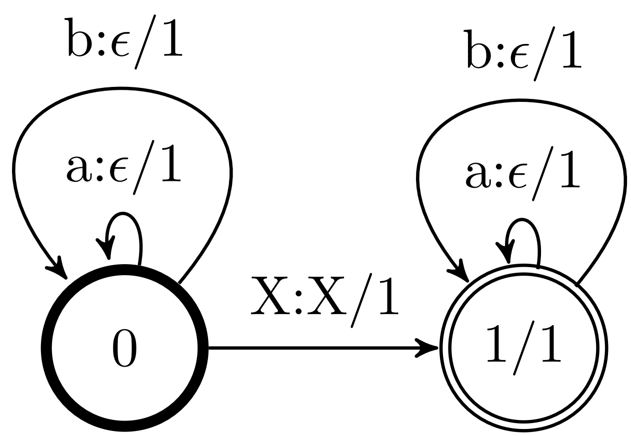


$$L(X) = \{aa, ab, ba, bb\}$$

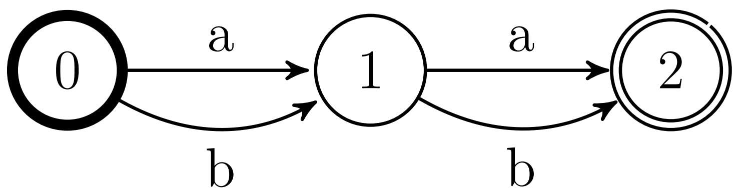
Figure 5.7: A Counting Transducer

- X : an automaton which generates a regular language $L(X)$ over the alphabet Σ.
- The "transition" $X : X/1$ stands for the part of the counting transducer created from the automaton X by adding to each transition an output label identical to the existing label, and by making all transition and final weights equal to 1.

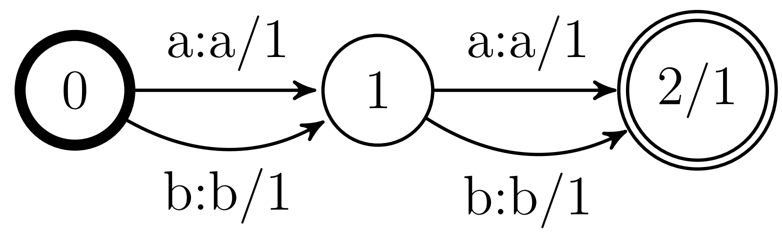
$\boxed{T_{\text{counting}} \text{ with } \Sigma = \{a, b\}}$



Example : Transformation of X to $X : X/1$



X : an automaton



$X : X/1$: a part of T_{count}

Constructing Counting Transducers from Automata

Theorem 5.10: Let

- Σ : a finite alphabet,
- X : a finite automaton over Σ,
- $L(X)$: the set of all strings in Σ^* accepted by the finite automaton X.

For any $x \in \Sigma^*$ and any sequence z accepted by an automaton X, i.e., $z \in L(X)$, $T_{\text{counting}}(x, z)$ is the number of occurrences of z in x.

Remarks

- The counting kernel $K_{\text{counting}} = T_{\text{counting}} \circ T_{\text{counting}}^{-1}$ is PDS.
- By changing the transition and/or final weights of the automaton X part in the definition of T_{count}, one can assign different weights to the patterns counted to emphasize or deemphasize some, as in the case of gappy_bigrams.