SC4/SM4 Data Mining and Machine Learning
Kernel Methods

Dino Sejdinovic
Department of Statistics
Oxford

Slides and other materials available at:
http://www.stats.ox.ac.uk/~sejdinov/dmml

Support vector classification

Regularised empirical risk minimisation problem with hinge loss.
Regularisation naturally arises from the margin penalty.

$$\min_{w,b} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \left(1 - y_i \left(w^\top x_i + b\right)\right)_+ \right).$$

Using substitution $\xi_i = \left(1 - y_i \left(w^\top x_i + b\right)\right)_+$, we obtain an equivalent formulation (primal C-SVM):

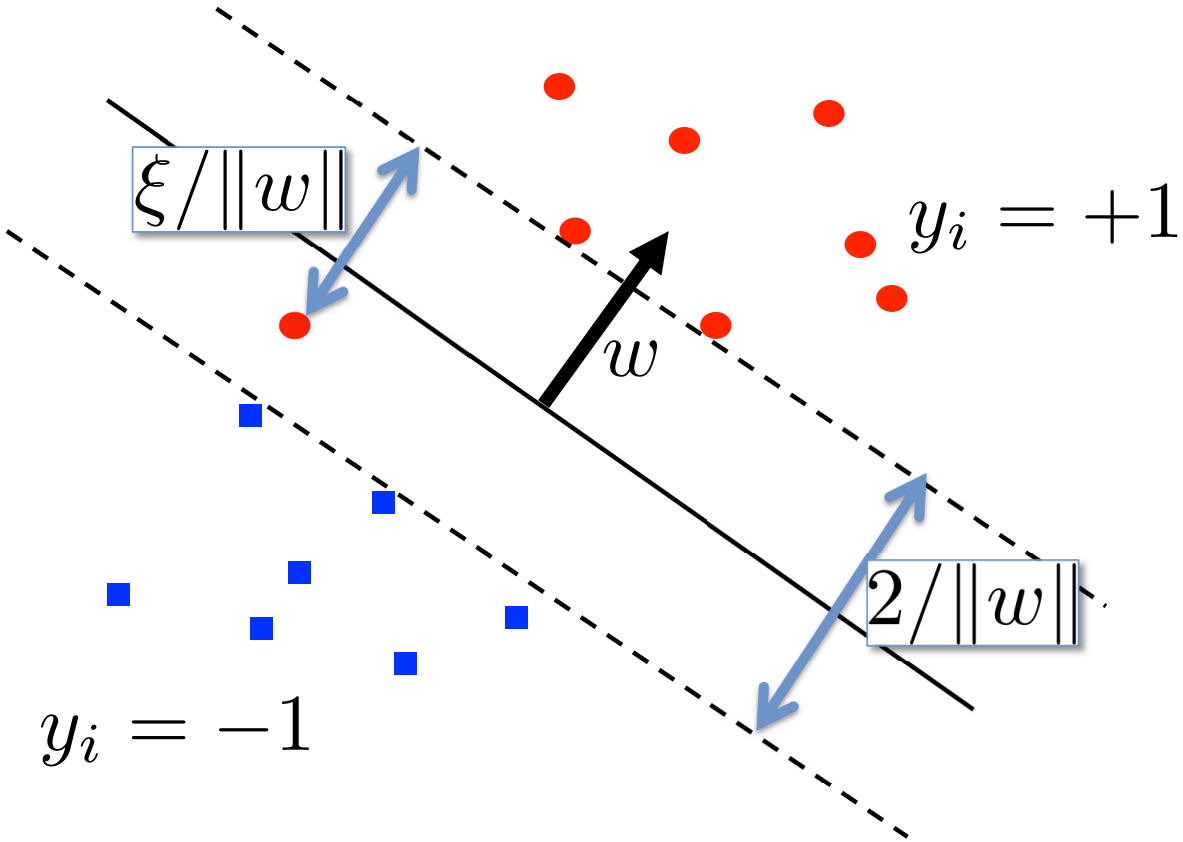
$$\min_{w,b,\xi} \left(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{n} \xi_i \right)$$

subject to

$$\xi_i \geq 0 \qquad y_i \left(w^\top x_i + b\right) \geq 1 - \xi_i$$

A convex constrained optimization problem with affine constraints in w, b, ξ: strong duality holds.

Support vector classification



Dual C-SVM

$$\text{maximize} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j,$$

subject to the constraints

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} y_i \alpha_i = 0$$

From α, obtain the hyperplane with

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i.$$

Offset b can be obtained from any of the margin SVs (for which $\alpha_i \in (0, C)$): $1 = y_i \left(w^\top x_i + b \right)$.

Dual form and Inner Products

We have stumbled across something quite interesting. Dual program

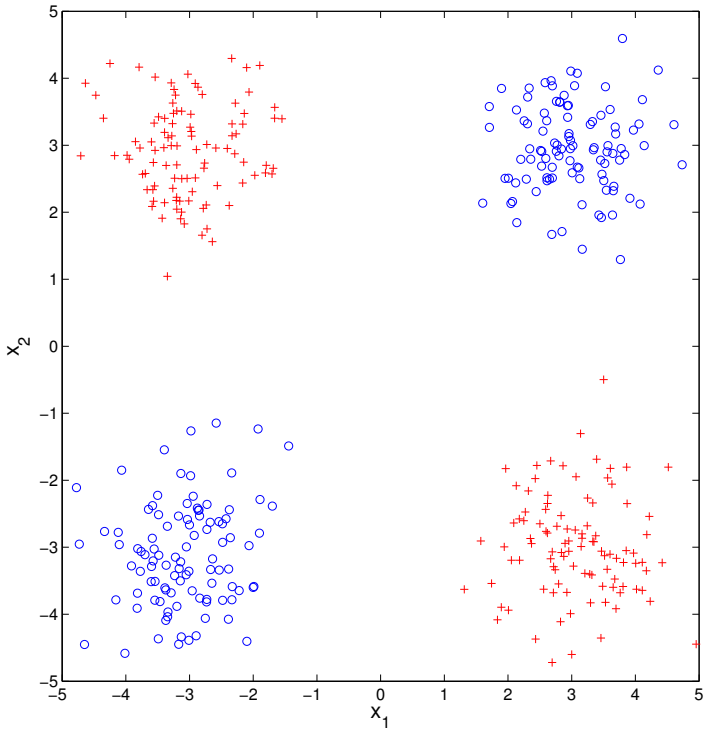
$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j \qquad \text{subject to} \quad \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \preceq \alpha \preceq C \end{cases}$$

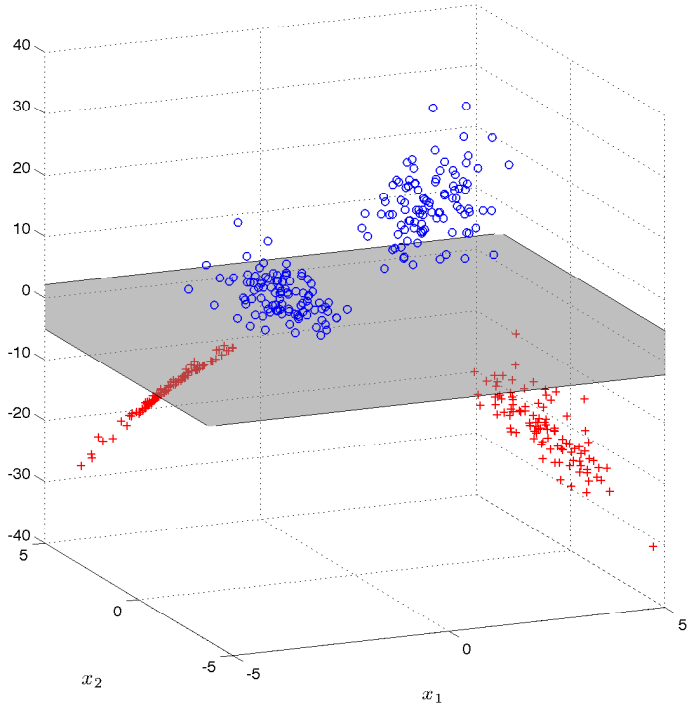
only depends on inputs x_i through their inner products (similarities) with other inputs.
Decision function

$$f(x) = \text{sign}(w^\top x + b) = \text{sign}(\sum_{i=1}^{n} \alpha_i y_i x_i^\top x + b)$$

also depends only on the similarity of a test point x to the training points x_i.
Thus, we do not need explicit inputs - just their pairwise similarities.
Key property: even if $p > n$, it is still the case that $w \in \text{span}\{x_i : i = 1, \ldots, n\}$ (normal vector of the hyperplane lives in the subspace spanned by the datapoints).

Beyond Linear Classifiers





- No linear classifier separates red from blue.
- Linear separation after mapping to a **higher dimensional feature space**:

$$\mathbb{R}^2 \ni \begin{pmatrix} x^{(1)} & x^{(2)} \end{pmatrix}^\top = x \mapsto \varphi(x) = \begin{pmatrix} x^{(1)} & x^{(2)} & x^{(1)}x^{(2)} \end{pmatrix}^\top \in \mathbb{R}^3$$

Non-Linear SVM

- Consider the dual C-SVM with explicit non-linear transformation $x \mapsto \varphi(x)$:

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \varphi(x_i)^\top \varphi(x_j) \qquad \text{subject to} \quad \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \preceq \alpha \preceq C \end{cases}$$

- Suppose $p = 2$, and we would like to introduce quadratic non-linearities,

$$\varphi(x) = \left(1, \sqrt{2}x^{(1)}, \sqrt{2}x^{(2)}, \sqrt{2}x^{(1)}x^{(2)}, \left(x^{(1)}\right)^2, \left(x^{(2)}\right)^2\right)^\top.$$

  Then

$$\varphi(x_i)^\top \varphi(x_j) = 1 + 2x_i^{(1)}x_j^{(1)} + 2x_i^{(2)}x_j^{(2)} + 2x_i^{(1)}x_i^{(2)}x_j^{(1)}x_j^{(2)} + \left(x_i^{(1)}\right)^2 \left(x_j^{(1)}\right)^2 + \left(x_i^{(2)}\right)^2 \left(x_j^{(2)}\right)^2 = (1 + x_i^\top x_j)^2$$

- Since only inner products are needed, non-linear transform need not be computed explicitly - inner product between features can be a simple function (**kernel**) of x_i and x_j: $k(x_i, x_j) = \varphi(x_i)^\top \varphi(x_j) = (1 + x_i^\top x_j)^2$
- d-order interactions can be implemented by $k(x_i, x_j) = (1 + x_i^\top x_j)^d$ (**polynomial kernel**). Never need to compute explicit feature expansion of dimension $\binom{p+d}{d}$ where this inner product happens!

Kernel SVM: Kernel trick

- Kernel SVM with $k(x_i, x_j)$. Non-linear transformation $x \mapsto \varphi(x)$ still present, but **implicit** (coordinates of the vector $\varphi(x)$ are never computed).

$$\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \qquad \text{subject to} \quad \begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ 0 \preceq \alpha \preceq C \end{cases}$$

- Prediction? $f(x) = \text{sign}\left(w^\top \varphi(x) + b\right)$, where $w = \sum_{i=1}^{n} \alpha_i y_i \varphi(x_i)$ and offset b obtained from a margin support vector x_j with $\alpha_j \in (0, C)$.
  - No need to compute w either! Just need

$$w^\top \varphi(x) = \sum_{i=1}^{n} \alpha_i y_i \varphi(x_i)^\top \varphi(x) = \sum_{i=1}^{n} \alpha_i y_i k(x_i, x).$$

  - Get offset from

$$b = y_j - w^\top \varphi(x_j) = y_j - \sum_{i=1}^{n} \alpha_i y_i k(x_i, x_j)$$

    for any margin support-vector x_j ($\alpha_j \in (0, C)$).
- Fitted a separating hyperplane in a high-dimensional feature space without ever mapping explicitly to that space.

Kernel trick in general

- In a learning algorithm, if only inner products $x_i^\top x_j$ are explicitly used, rather than data items x_i, x_j directly, we can replace them with a kernel function $k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$, where $\varphi(x)$ could be **nonlinear, high- and potentially infinite-dimensional** features of the original data.
  - Kernel ridge regression
  - Kernel logistic regression
  - Kernel PCA, CCA, ICA
  - Kernel K-means

Kernel Methods and Reproducing Kernel Hilbert Spaces

slides based on Arthur Gretton's Reproducing kernel Hilbert spaces in Machine Learning course

Kernel: an inner product between feature maps

Definition (kernel)

Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if there exists a **Hilbert space** and a map $\varphi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$k(x, x') := \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}} .$$

- Almost no conditions on $\mathcal{X}$ (eg, $\mathcal{X}$ itself need not have an inner product, e.g., documents).
- Think of kernel as a **similarity measure between features**

What are some simple kernels? E.g., for text documents? For images?

- A single kernel can correspond to multiple sets of underlying features.

$$\varphi_1(x) = x \qquad \text{and} \qquad \varphi_2(x) = \begin{pmatrix} x/\sqrt{2} & x/\sqrt{2} \end{pmatrix}^\top$$

Positive semidefinite functions

If we are given a "measure of similarity" with two arguments, $k(x, x')$, how can we determine if it is a valid kernel?

1. Find a feature map?
   - Sometimes not obvious (especially if the feature vector is infinite dimensional)
2. A simpler direct property of the function: positive semidefiniteness.

Positive semidefinite functions

Definition (Positive semidefinite functions)

A symmetric function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semidefinite if $\forall n \geq 1,\ \forall (a_1, \ldots a_n) \in \mathbb{R}^n,\ \forall (x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \kappa(x_i, x_j) \geq 0.$$

- Kernel $k(x, y) := \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$ for a Hilbert space $\mathcal{H}$ is positive semidefinite.

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \varphi(x_i), a_j \varphi(x_j) \rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^{n} a_i \varphi(x_i) \right\|_{\mathcal{H}}^2 \geq 0. \end{aligned}$$

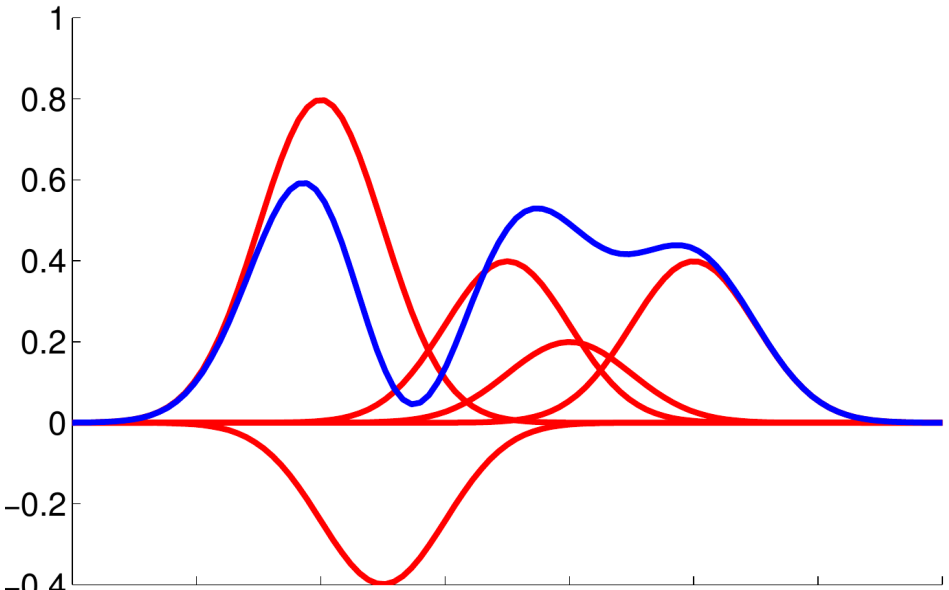
Positive semidefinite functions are kernels

Moore-Aronszajn Theorem

Every positive semidefinite function is a kernel for some Hilbert space $\mathcal{H}$.

- $\mathcal{H}$ is usually thought of as a space of functions (**Reproducing kernel Hilbert space - RKHS**)

Gaussian RBF kernel $k(x, x') = \exp\left(-\frac{1}{2\gamma^2}\|x - x'\|^2\right)$ has an infinite-dimensional $\mathcal{H}$ with elements $h(x) = \sum_{i=1}^{m} a_i k(x_i, x)$ (recall that $w^\top \varphi(x)$ in SVM has exactly this form!).



Reproducing kernel

Definition (Reproducing kernel)

Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ defined on a non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called **a reproducing kernel** of $\mathcal{H}$ if it satisfies

- $\forall x \in \mathcal{X}, \quad k_x = k(\cdot, x) \in \mathcal{H}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (the reproducing property).

In particular, for any $x, y \in \mathcal{X}$, $k(x, y) = \langle k(\cdot, y), k(\cdot, x) \rangle_{\mathcal{H}} = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}$.

Can forget all about $\varphi(x)$ and just treat $k(\cdot, x)$ as a feature of x (it is a perfectly valid Hilbert-space valued feature)!

RKHS

Definition (Reproducing kernel Hilbert space)

A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$ is said to be a Reproducing Kernel Hilbert Space (RKHS) if evaluation functionals $\delta_x : \mathcal{H} \to \mathbb{R}$, $\delta_x f = f(x)$ are continuous $\forall x \in \mathcal{X}$.

Theorem (Norm convergence implies pointwise convergence)

If $\lim_{n\to\infty} \|f_n - f\|_{\mathcal{H}} = 0$, *then* $\lim_{n\to\infty} f_n(x) = f(x)$, $\forall x \in \mathcal{X}$.

- If two functions $f, g \in \mathcal{H}$ are close in the norm of $\mathcal{H}$, then $f(x)$ and $g(x)$ are close for all $x \in \mathcal{X}$
- This is a property of particularly "nice" functional spaces. For example, does not hold on spaces endowed with L_2 norm: x^n on $[0, 1]$ converges to 0 in L_2 but not pointwise.

Back to SVMs

Maximum margin classifier in RKHS: Looking for a decision function of form $\text{sign}(w(x))$ where $w \in \mathcal{H}_k$. Because we are in an RKHS, $w = \langle w, k(\cdot, x)\rangle_{\mathcal{H}_k}$.

$$\min_{w \in \mathcal{H}_k} \left(\frac{1}{2}\|w\|^2_{\mathcal{H}_k} + C \sum_{i=1}^{n} h\left(y_i \langle w, k(\cdot, x_i)\rangle_{\mathcal{H}_k}\right) \right)$$

for the RKHS $\mathcal{H}$ with kernel $k(x, x')$. Maximizing the margin equivalent to minimizing $\|w\|^2_{\mathcal{H}}$: for many RKHSs a smoothness constraint on function w (more about this later).
Why can we solve this infinite-dimensional optimization problem? Because we know that $w \in \text{span}\{k(\cdot, x_i) : i = 1, \ldots, n\}$ – Representer Theorem.

Representer Theorem

Representer theorem

Standard supervised learning setup: we are given a set of paired observations $(x_1, y_1), \ldots (x_n, y_n)$.
Goal: find the function f^* in the RKHS $\mathcal{H}$ which solves the regularized empirical risk minimization problem.

$$\min_{f \in \mathcal{H}} \hat{R}(f) + \Omega\left(\|f\|_{\mathcal{H}}^2\right),$$

where empirical risk is

$$\hat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i), x_i),$$

and Ω is a non-decreasing function.

- Classification: L could be a hinge loss $L(y, f(x), x) = (1 - yf(x))_+$ or a logistic loss $L(y, f(x), x) = \log(1 + \exp(-yf(x))$.
- Regression: $L(y, f(x), x) = (y - f(x))^2$.

Representer theorem

Theorem (Representer Theorem)

There is a solution to

$$\min_{f \in \mathcal{H}} \hat{R}(f) + \Omega\left(\|f\|_{\mathcal{H}}^2\right)$$

that takes the form

$$f^* = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i).$$

If Ω is strictly increasing, all solutions have this form.

Representer theorem: proof

Proof: Denote f_s projection of f onto the subspace

$$\text{span}\left\{k(\cdot, x_i) : \ i = 1, \ldots, n\right\}$$

such that

$$f = f_s + f_\perp,$$

where $f_s = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ and $f_\perp$ is orthogonal to $\text{span}\left\{k(\cdot, x_i) : \ i = 1, \ldots, n\right\}$.
Regularizer:

$$\|f\|_{\mathcal{H}}^2 = \|f_s\|_{\mathcal{H}}^2 + \|f_\perp\|_{\mathcal{H}}^2 \geq \|f_s\|_{\mathcal{H}}^2 ,$$

then

$$\Omega\left(\|f\|_{\mathcal{H}}^2\right) \geq \Omega\left(\|f_s\|_{\mathcal{H}}^2\right).$$

Representer theorem: proof

Proof (cont.): Individual terms $f(x_i)$ in the loss:

$$f(x_i) = \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}} = \langle f_s + f_\perp, k(\cdot, x_i)\rangle_{\mathcal{H}} = \langle f_s, k(\cdot, x_i)\rangle_{\mathcal{H}} ,$$

so

$$L(y_i, f(x_i), x_i) = L(y_i, f_s(x_i), x_i) \forall i \implies \hat{R}(f) = \hat{R}(f_s).$$

Hence

- The empirical risk only depends on the components of f lying in the subspace spanned by canonical features.
- Regularizer $\Omega(\ldots)$ is minimized when $f = f_s$.
- If Ω is strictly non-decreasing, then $\|f_\perp\|_{\mathcal{H}} = 0$ is required at the minimum.

Kernel Ridge Regression

Regularised Least Squares

We are given n training points $\{x_i\}_{i=1}^n$ in $\mathbb{R}^p$: Define some $\lambda > 0$. Our goal is:

$$\begin{aligned} w^* &= \arg\min_{w \in \mathbb{R}^p} \left(\sum_{i=1}^n (y_i - x_i^\top w)^2 + \lambda \|w\|^2 \right) \\ &= \arg\min_{w \in \mathbb{R}^p} \left(\|\mathbf{y} - \mathbf{X}w\|^2 + \lambda \|w\|^2 \right), \end{aligned}$$

Solution is:

$$w^* = \left(\mathbf{X}^\top \mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^\top \mathbf{y},$$

which is the standard regularised least squares solution.

Kernel ridge regression

Use features $\phi(x_i)$ in the place of x_i:

$$w^* \quad = \quad \arg\min_{w \in \mathcal{H}} \left(\sum_{i=1}^{n} (y_i - \langle w, \phi(x_i) \rangle_{\mathcal{H}})^2 + \lambda \|w\|_{\mathcal{H}}^2 \right).$$

E.g. for finite dimensional feature spaces,

$$\phi_p(x) = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^\ell \end{bmatrix} \qquad \phi_s(x) = \begin{bmatrix} \sin(x) \\ \cos(x) \\ \sin(2x) \\ \vdots \\ \cos\left(\frac{\ell}{2}x\right) \end{bmatrix}$$

In finite dimensions, w is a vector of length ℓ giving weight to each of these features so that learned function is $f_w(x) = w^\top \phi(x)$. Feature vectors can also have **infinite** length.

Kernel ridge regression

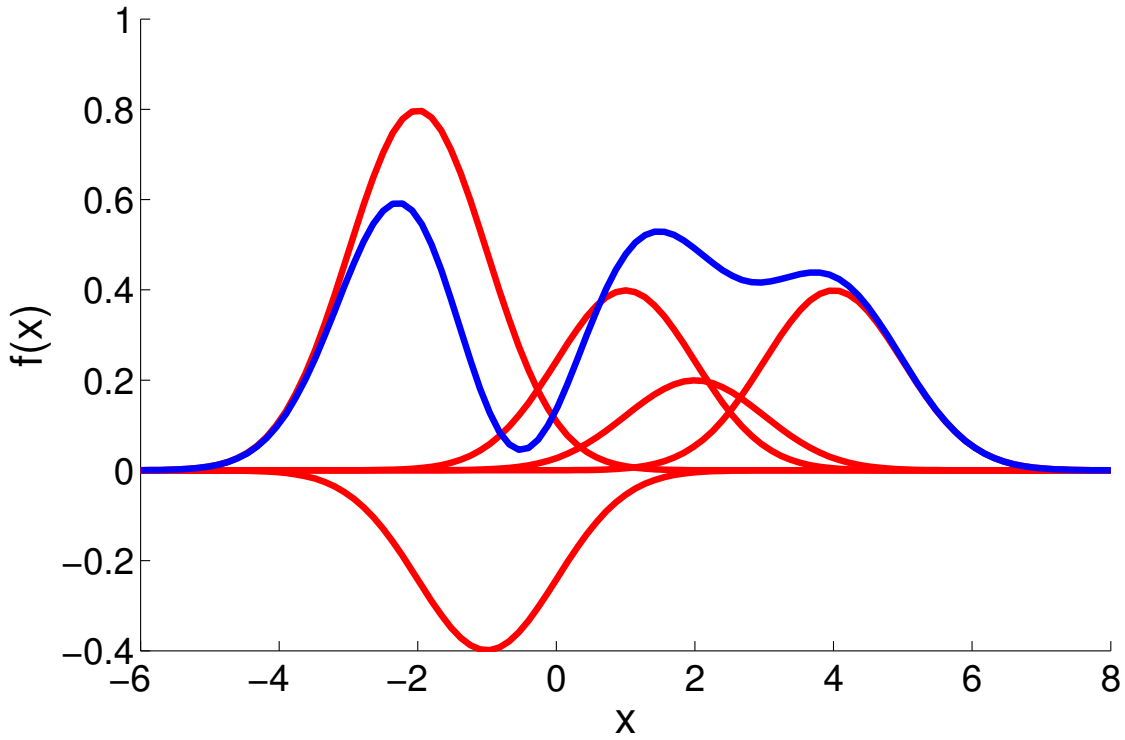
Recall that feature maps ϕ and feature spaces $\mathcal{H}$ are not unique, but RKHS $\mathcal{H}_k$ is. Thus, we can identify w with the function f_w (there is an isometry between w and f_w: $\|w\|_{\mathcal{H}} = \|f_w\|_{\mathcal{H}_k}$ regardless of the choice of the feature space $\mathcal{H}$) and write

$$\begin{aligned} f^* &= \arg\min_{f \in \mathcal{H}_k} \left(\sum_{i=1}^{n} (y_i - \langle f, k(\cdot, x_i) \rangle_{\mathcal{H}})^2 + \lambda \|f\|_{\mathcal{H}_k}^2 \right) \\ &= \arg\min_{f \in \mathcal{H}_k} \left(\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}_k}^2 \right). \end{aligned}$$

Kernel ridge regression

Recall the representer theorem: f is a linear combination of feature space mappings of data points

$$w = \sum_{i=1}^{n} \alpha_i \phi(x_i), \quad f_w = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot).$$



Kernel ridge regression

Recall the representer theorem: f is a linear combination of feature space mappings of data points

$$f = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i).$$

Then

$$\begin{aligned} \sum_{i=1}^{n} \left(y_i - \langle f, k(\cdot, x_i)\rangle_{\mathcal{H}_k}\right)^2 + \lambda \|f\|_{\mathcal{H}_k}^2 &= \|\mathbf{y} - \mathbf{K}\alpha\|^2 + \lambda \alpha^\top \mathbf{K}\alpha \\ &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top K\alpha + \alpha^\top \left(\mathbf{K}^2 + \lambda \mathbf{K}\right)\alpha \end{aligned}$$

Differentiating wrt α and setting this to zero, we get

$$\alpha^* = (\mathbf{K} + \lambda I_n)^{-1} y.$$

Recall: $\frac{\partial \alpha^\top U \alpha}{\partial \alpha} = (U + U^\top)\alpha, \qquad \frac{\partial v^\top \alpha}{\partial \alpha} = \frac{\partial \alpha^\top v}{\partial \alpha} = v$

Parameter selection for KRR

Given the objective

$$f^* = \arg\min_{f \in \mathcal{H}_k} \left(\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|^2_{\mathcal{H}_k} \right).$$

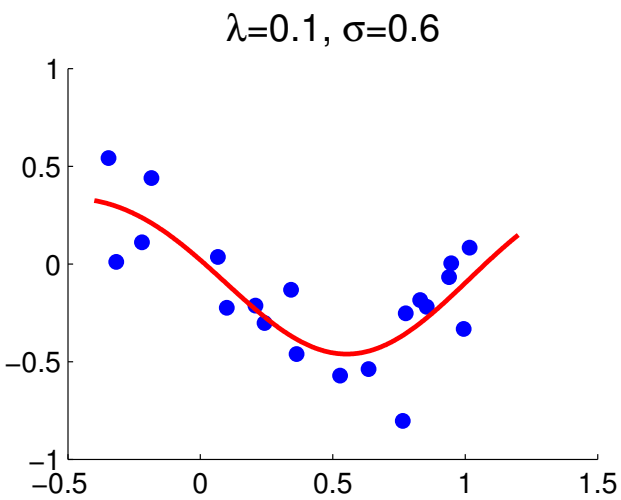
How do we choose

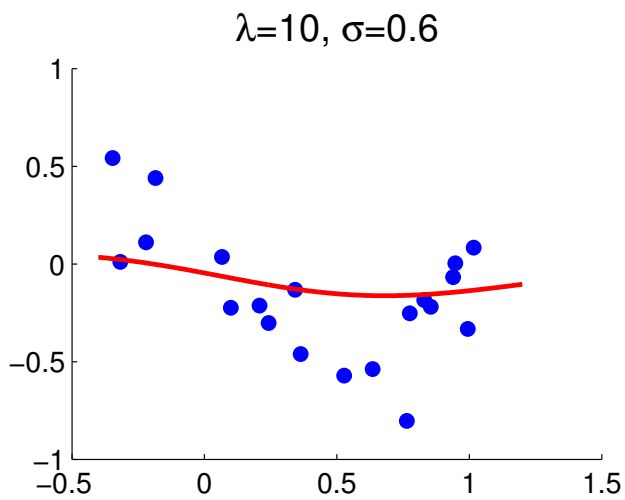
- The regularization parameter λ?
- The kernel parameter: for Gaussian kernel, σ in

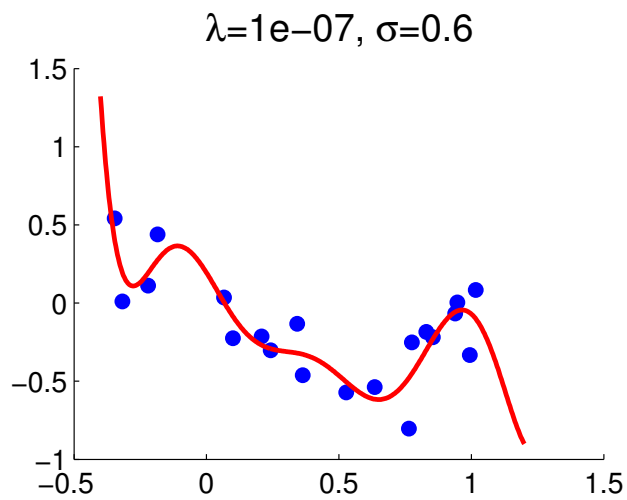
$$k(x, y) = \exp\left(\frac{-\|x - y\|^2}{\sigma} \right).$$

Beware: Gaussian kernel has many different parametrisations in the literature and software packages!
Typically use cross-validation.

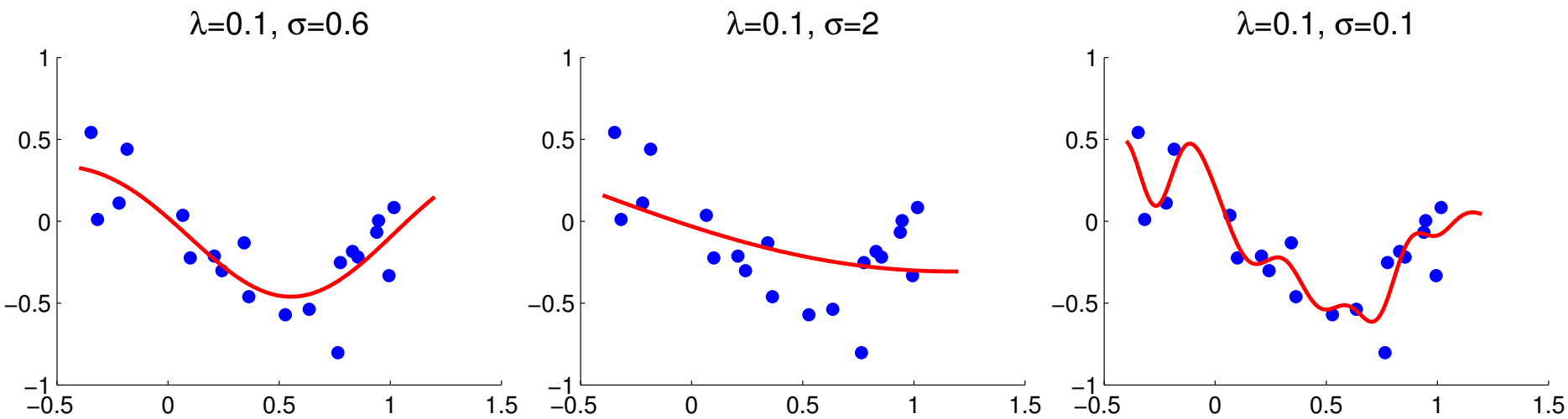
Choice of λ







Choice of σ



Kernel Assembly Line

Examples of kernels

- **Linear**: $k(x, x') = x^\top x'$.
- **Polynomial**: $k(x, x') = (c + x^\top x')^m$, $c \in \mathbb{R}$, $m \in \mathbb{N}$.
- **Exponential**: $k(x, x') = \exp(\frac{x^\top x'}{\gamma})$, $\gamma > 0$.
- **Gaussian RBF**: $k(x, x') = \exp\left(-\frac{1}{2\gamma^2}\|x - x'\|^2\right)$, $\gamma > 0$.
- **Laplacian**: $k(x, x') = \exp\left(-\frac{1}{2\gamma^2}\|x - x'\|\right)$, $\gamma > 0$.
- **Rational quadratic**: $k(x, x') = \left(1 + \frac{\|x-x'\|^2}{2\alpha\gamma^2}\right)^{-\alpha}$, $\alpha, \gamma > 0$.
- **Brownian covariance**: $k(x, x') = \frac{1}{2}\left(\|x\|^\gamma + \|x'\|^\gamma - \|x - x'\|^\gamma\right)$, $\gamma \in [0, 2]$.

New kernels from old: sums, transformations

The great majority of useful kernels are built from simpler kernels.

Lemma (Sums of kernels are kernels)

Given $\alpha > 0$ and k, k_1 and k_2 all kernels on $\mathcal{X}$, then αk and $k_1 + k_2$ are kernels on $\mathcal{X}$.

To prove this, just check inner product definition (features get scaled with $\sqrt{\alpha}$ or concatenated). A difference of kernels need not be a kernel (**why?**)

Lemma (Space transformation)

Let $\mathcal{X}$ and $\widetilde{\mathcal{X}}$ be sets, and consider any map $s : \mathcal{X} \to \widetilde{\mathcal{X}}$. Let $\tilde{k}$ be a kernel on $\widetilde{\mathcal{X}}$. Then $k(x, x') = \tilde{k}(s(x), s(x'))$ is a kernel on $\mathcal{X}$.

Proof: if $\tilde{\varphi}$ is a feature map for $\tilde{k}$, then $\varphi = \tilde{\varphi} \circ s$ is a feature map for k.

New kernels from old: products

Lemma (Products of kernels are kernels)

Given k_1 on $\mathcal{X}_1$ and k_2 on $\mathcal{X}_2$, then $k_1 \times k_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$.

Proof.

Sketch for finite-dimensional spaces only. Assume $\mathcal{H}_1$ corresponding to k_1 is $\mathbb{R}^m$, and $\mathcal{H}_2$ corresponding to k_2 is $\mathbb{R}^n$. Define:

- $k_1 := u^\top v$ for $u, v \in \mathbb{R}^m$ (e.g.: kernel between two images)
- $k_2 := p^\top q$ for $p, q \in \mathbb{R}^n$ (e.g.: kernel between two captions)

Is the following a kernel?

$$K\left[(u,p);(v,q)\right] = k_1 \times k_2$$

(e.g. kernel between one image-caption pair and another)

□

New kernels from old: products

Proof.

(continued)

$$\begin{aligned} k_1 k_2 &= \left(u^\top v\right)\left(q^\top p\right) \\ &= \operatorname{trace}(u^\top v q^\top p) \\ &= \operatorname{trace}(p u^\top v q^\top) \\ &= \langle A, B \rangle , \end{aligned}$$

where $A := pu^\top$, $B := qv^\top$ (features of image-caption pairs) Thus $k_1 k_2$ is a valid kernel, since inner product between $A, B \in \mathbb{R}^{m \times n}$ is

$$\langle A, B \rangle = \operatorname{trace}(AB^\top).$$

□

Another way: just note that the **Kronecker product of positive definite matrices is positive definite**!

More products and Taylor expansions

Lemma (Products of kernels are kernels)

Given kernels k_1 and k_2 on $\mathcal{X}$ $k_1 \times k_2$ is a kernel on $\mathcal{X}$.

Proof: It is certainly a kernel on $\mathcal{X} \times \mathcal{X}$, so just consider space transformation $s : \mathcal{X} \to \mathcal{X} \times \mathcal{X}$ with $s(x) = (x, x)$.
Another way: just note that the **Hadamard product of positive definite matrices is positive definite**!
As a corollary:

$$k(x, x') = c + \sum_{j=1}^{d} a_j \langle x, x' \rangle^d \tag{1}$$

is certainly a kernel. Readily extends to

$$k(x, x') = g\left(\langle x, x' \rangle\right) \tag{2}$$

for an analytic function g with nonnegative Taylor coefficients, e.g., $\exp$.

Gaussian RBF is a kernel

As a product of an exponential kernel and a kernel with 1-d feature $x \mapsto \exp\left(-\frac{\|x\|^2}{2\gamma^2}\right)$.

$$\begin{aligned} k(x, x') &= \exp\left(-\frac{1}{2\gamma^2}\|x - x'\|^2\right) \\ &= \exp\left(-\frac{\|x\|^2}{2\gamma^2}\right)\exp\left(-\frac{\|x'\|^2}{2\gamma^2}\right)\exp\left(\frac{1}{\gamma^2}\langle x, x'\rangle\right) \end{aligned}$$

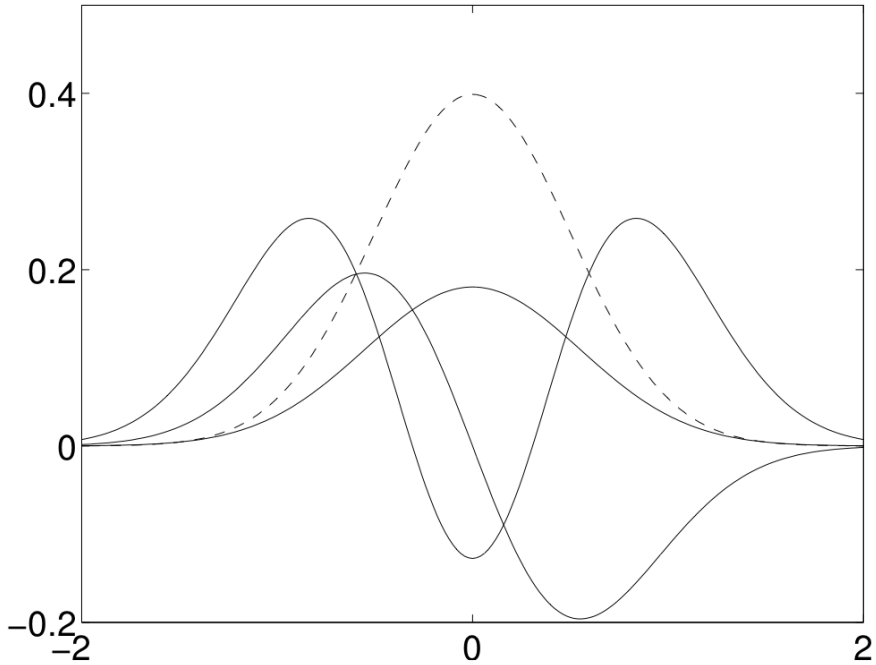
All of the proofs above are constructive: they give a way of constructing new features from old. But the resulting features quickly become very difficult to interpret. There is another, much cleaner way to do this: Mercer's Theorem.

Mercer's Theorem and Smoothness

What does $\|f\|_{\mathcal{H}}$ have to do with smoothing? For the Gaussian kernel:

$$f(x) = \sum_{r=1}^{\infty} a_r e_r(x), \qquad \|f\|_{\mathcal{H}}^2 = \sum_{r=1}^{\infty} \frac{a_r^2}{\lambda_r}.$$

$\lambda_r \sim B^r \to 0$, as $r \to \infty$ for $B \in (0,1)$ and $e_r(x)$ are functions of increasing complexity as r increases (r zero-crossings) – related to r-th order **Hermite polynomials**. Figure from Rasmussen and Williams, 2006



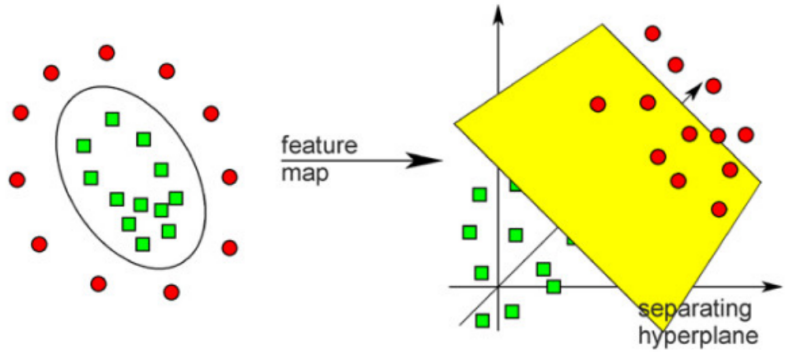
Examples of kernels

- **Linear**: $k(x, x') = x^\top x'$.
- **Polynomial**: $k(x, x') = (c + x^\top x')^m$, $c \in \mathbb{R}$, $m \in \mathbb{N}$.
- **Exponential**: $k(x, x') = \exp(\frac{x^\top x'}{\gamma})$, $\gamma > 0$.
- **Gaussian RBF**: $k(x, x') = \exp\left(-\frac{1}{2\gamma^2}\|x - x'\|^2\right)$, $\gamma > 0$.
- **Laplacian**: $k(x, x') = \exp\left(-\frac{1}{2\gamma^2}\|x - x'\|\right)$, $\gamma > 0$.
- **Rational quadratic**: $k(x, x') = \left(1 + \frac{\|x-x'\|^2}{2\alpha\gamma^2}\right)^{-\alpha}$, $\alpha, \gamma > 0$.
- **Brownian covariance**: $k(x, x') = \frac{1}{2}\left(\|x\|^\gamma + \|x'\|^\gamma - \|x - x'\|^\gamma\right)$, $\gamma \in [0, 2]$.

RKHS Embeddings of Distributions

Kernel Trick and Kernel Mean Trick

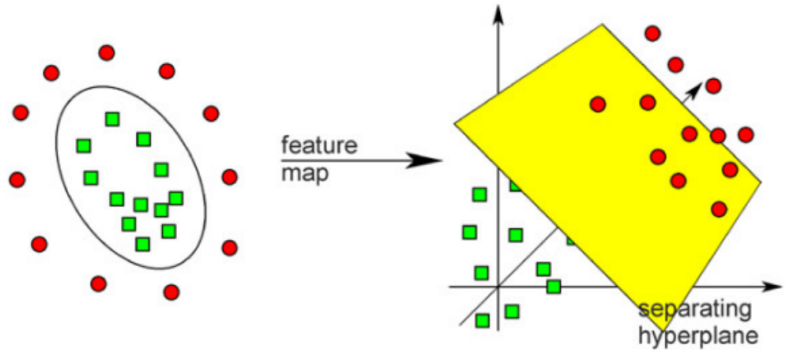
- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\varphi_1(x), \ldots, \varphi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data



[Cortes & Vapnik, 1995; Schölkopf & Smola, 2001]

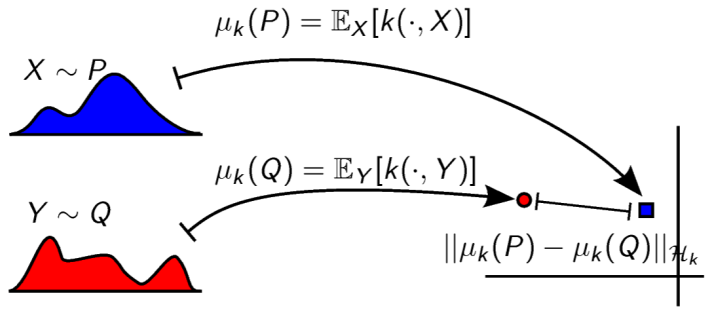
Kernel Trick and Kernel Mean Trick

- implicit feature map $x \mapsto k(\cdot, x) \in \mathcal{H}_k$
  replaces $x \mapsto [\varphi_1(x), \ldots, \varphi_s(x)] \in \mathbb{R}^s$
- $\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_k} = k(x, y)$
  inner products readily available
  - nonlinear decision boundaries, nonlinear regression functions, learning on non-Euclidean/structured data



[Cortes & Vapnik, 1995; Schölkopf & Smola, 2001]

- **RKHS embedding**: implicit feature mean
  [Smola et al, 2007; Sriperumbudur et al, 2010]
  $P \mapsto \mu_k(P) = \mathbb{E}_{X \sim P} k(\cdot, X) \in \mathcal{H}_k$
  replaces $P \mapsto [\mathbb{E}\varphi_1(X), \ldots, \mathbb{E}\varphi_s(X)] \in \mathbb{R}^s$
- $\langle \mu_k(P), \mu_k(Q) \rangle_{\mathcal{H}_k} = \mathbb{E}_{X \sim P, Y \sim Q} k(X, Y)$
  inner products easy to estimate
  - multiple instance learning / learning on distributions, nonparametric testing for homogeneity, independence, conditional independence, three-variable interaction



[Gretton et al, 2005; Gretton et al, 2006; Fukumizu et al, 2007; DS, Bergsma & Gretton, 2013; Szabo et al, 2015]

Learning on Distributions

- Multiple-Instance Learning: Input is a bag of B_i vectors $\{x_{i1}, \ldots, x_{iB_i}\}$, each $x_{ia} \in X$ assumed to arise from a probability distribution P_i on $\mathcal{X}$.
- Represent the i-th bag by the corresponding empirical kernel embedding $\mathfrak{m}_i = \mu_k[\mathsf{P}_i] = \frac{1}{B_i}\sum_{a=1}^{B_i} k(\cdot, x_{ia})$ w.r.t. a kernel k on $\mathcal{X}$.
- Now treat the problem as having inputs $\mathfrak{m}_i \in \mathcal{H}_k$: just need to define a **kernel** K on $\mathcal{H}_k$.

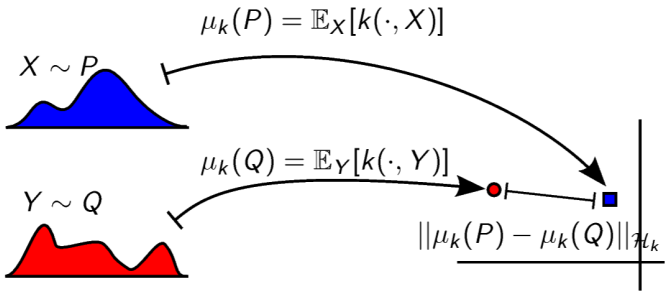
$$\text{Linear:} \qquad K(\mathfrak{m}_i, \mathfrak{m}_j) = \langle \mathfrak{m}_i, \mathfrak{m}_j \rangle_{\mathcal{H}_k} = \frac{1}{B_i B_j} \sum_{a=1}^{B_i} \sum_{b=1}^{B_j} k(x_{ia}, x_{jb})$$

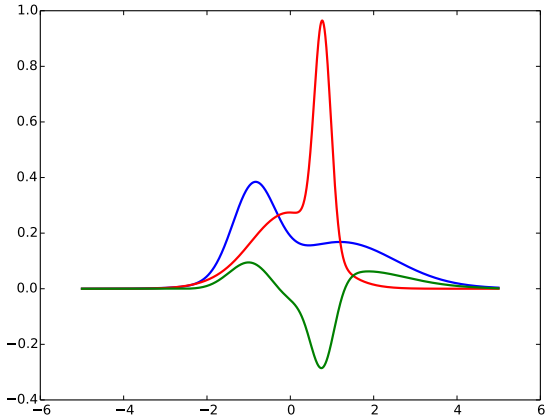
$$\text{Gaussian:} \qquad K(\mathfrak{m}_i, \mathfrak{m}_j) = \exp\left(-\frac{1}{2\gamma^2} \|\mathfrak{m}_i - \mathfrak{m}_j\|^2_{\mathcal{H}_k}\right).$$

Term $\|\mathfrak{m}_i - \mathfrak{m}_j\|^2_{\mathcal{H}_k}$ can be thought of as a distance between empirical measures corresponding to bags i and j. This is called Maximum Mean Discrepancy (MMD).

Maximum Mean Discrepancy

- **Maximum Mean Discrepancy (MMD)** [Borgwardt et al, 2006; Gretton et al, 2007] between P and Q:





$$\mathrm{MMD}_k(P, Q) = \|\mu_k(P) - \mu_k(Q)\|_{\mathcal{H}_k} = \sup_{f \in \mathcal{H}_k:\, \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E}f(X) - \mathbb{E}f(Y)|$$

- **Characteristic** kernels: $\mathrm{MMD}_k(P, Q) = 0$ iff $P = Q$.
  - Gaussian RBF, Matérn family, inverse multiquadrics...
  - For characteristic kernels on LCH $\mathcal{X}$, MMD metrizes weak* topology on probability measures [Sriperumbudur,2010],

$$\mathrm{MMD}_k(P_n, P) \to 0 \Leftrightarrow P_n \rightsquigarrow P.$$

- Kernel embedding represents expectations of RKHS functions:

$$\langle f, \mu_k[P] \rangle_{\mathcal{H}_k} = \int f(x) P(dx).$$

Two-sample testing on nonstandard domains

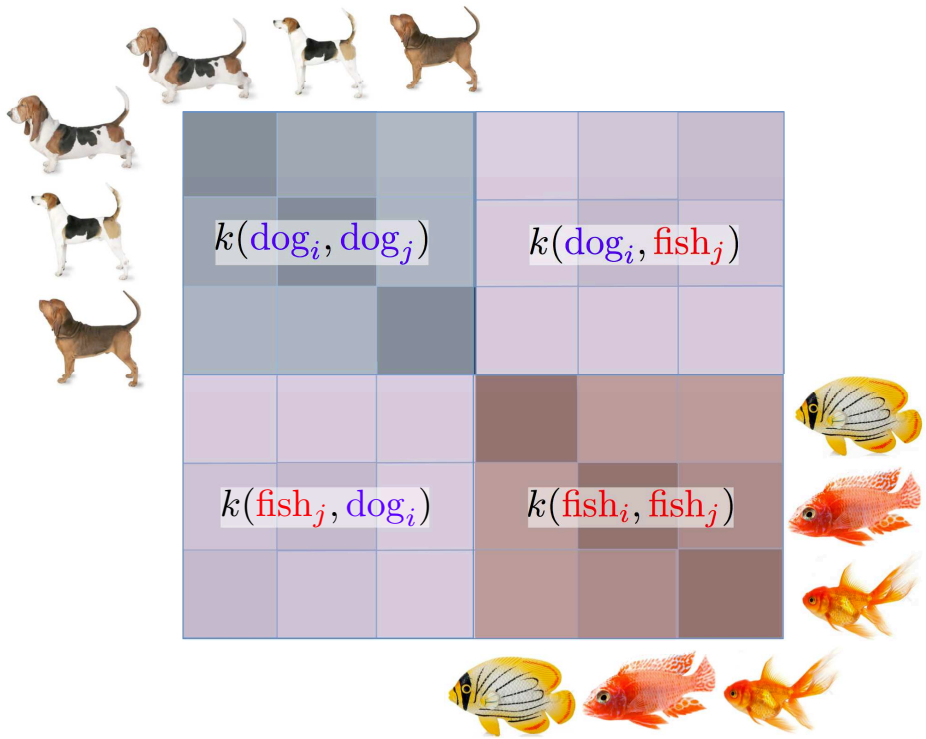


Figure by Arthur Gretton

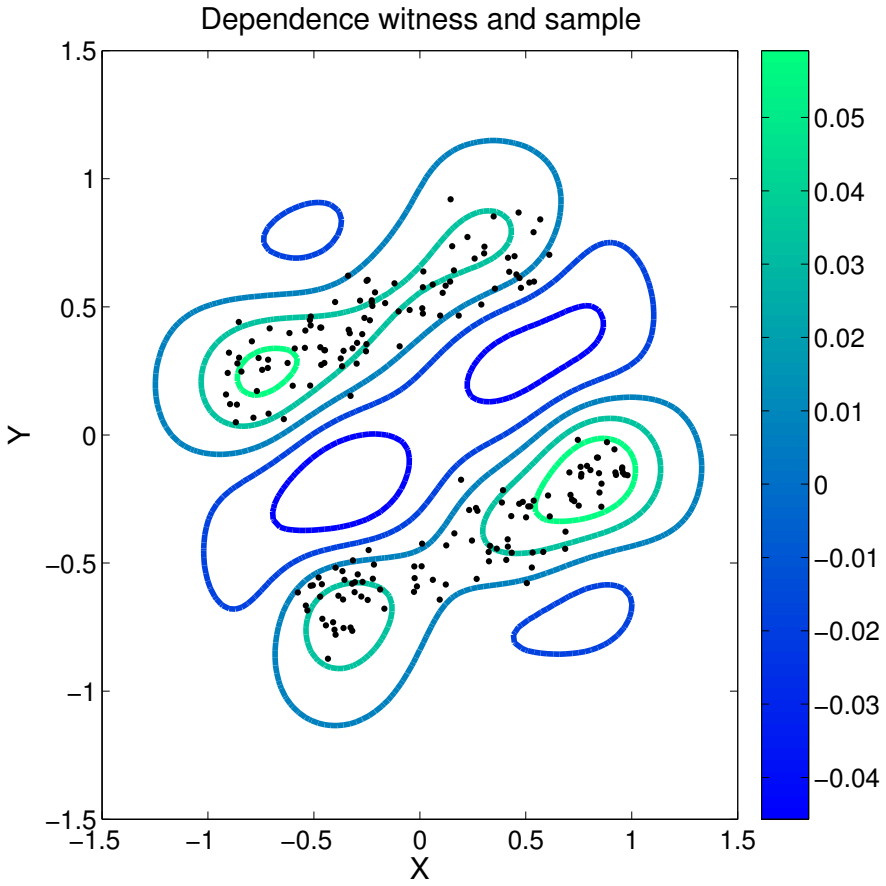
Average similarity within two samples vs average similarity across two samples.

MMD has been applied to:

- independence tests on text data [Gretton et al, 2009]
- two-sample tests on graphs [Gretton et al, 2012]
- training generative neural networks for image data [Dziugaite, Roy and Ghahramani, 2015]
- two-sample tests on persistence diagrams in topological data analysis [Kwitt et al, 2015]
- similarity measure between observed and simulated data in ABC [Park, Jitkrittum and DS, 2015]

$$\text{MMD}_k^2(P, Q) = \mathbb{E}_{X,X' \overset{i.i.d.}{\sim} P} k(X, X') + \mathbb{E}_{Y,Y' \overset{i.i.d.}{\sim} Q} k(Y, Y') - 2\mathbb{E}_{X \sim P, Y \sim Q} k(X, Y).$$

Kernel dependence measures



- $HSIC^2(X, Y; \kappa) = \|\mu_\kappa(P_{XY}) - \mu_\kappa(P_X P_Y)\|^2_{\mathcal{H}_\kappa}$
- dependence witness is a smooth function in the RKHS $\mathcal{H}_\kappa$ of functions on $\mathcal{X} \times \mathcal{Y}$

$k(\boxed{1}, \boxed{2})$ $\quad l(\boxed{1}, \boxed{2})$

$\kappa(\boxed{1}\boxed{1}, \boxed{2}\boxed{2}) =$

$k(\boxed{1}, \boxed{2}) \times l(\boxed{1}, \boxed{2})$

- Independence testing framework that generalises Distance Covariance (dCov): HSIC with Brownian motion covariance kernels

[Szekely et al, 2009; DS et al, 2013]

Large Scale Approximations

Kernel methods at scale

- Expressivity of kernel methods (rich, often infinite-dimensional hypothesis spaces) comes with a cost that scales at least quadratically in the number of observations n (due to needing to compute, store and often invert the Gram matrix)! We arrived at this by trying to avoid paying the cost in the dimension of the hypothesis space (e.g., for order d polynomial kernels, scales as $\binom{p+d}{d}$, and infinite for many kernels).
- But now we have to pay in terms of n which is problematic when we have a lot of observations (and this is exactly when we want to use a rich expressive model with a high-dimensional hypothesis class!)
- Scaling up kernel methods is a very active research area

  [Sonnenburg et al, 2006; Rahimi & Recht 2007; Le, Sarlos & Smola, 2013; Wilson et al, 2014; Dai et al, 2014; Sriperumbudur & Szabo, 2015].
- Main idea: study the desired hypothesis space and scale its dimension down - then undo the kernel trick!
- Errm... So we went the full circle (!?)
  explicit basis functions $\rightarrow$ implicit basis functions $\rightarrow$ explicit basis functions

Random Fourier features: Inverse Kernel Trick

Bochner's representation: any positive definite **translation-invariant** kernel on $\mathbb{R}^p$ can be written as

$$\begin{aligned} k(x,y) &= \int_{\mathbb{R}^p} \exp\left(i\omega^\top (x-y)\right) d\Lambda(\omega) \\ &= \int_{\mathbb{R}^p} \left\{\cos\left(\omega^\top x\right)\cos\left(\omega^\top y\right) + \sin\left(\omega^\top x\right)\sin\left(\omega^\top y\right)\right\} d\Lambda(\omega) \end{aligned}$$

for some positive measure (w.l.o.g. a probability distribution) Λ.

- Sample m frequencies $\{\omega_j\} \sim \Lambda$ and use a Monte Carlo estimator of the kernel function instead [Rahimi & Recht, 2007]:

$$\begin{aligned} \hat{k}(x,y) &= \frac{1}{m}\sum_{j=1}^{m} \left\{\cos\left(\omega_j^\top x\right)\cos\left(\omega_j^\top y\right) + \sin\left(\omega_j^\top x\right)\sin\left(\omega_j^\top y\right)\right\} \\ &= \langle \varphi_\omega(x), \varphi_\omega(y)\rangle_{\mathbb{R}^{2m}}, \end{aligned}$$

  with an explicit set of features $x \mapsto \frac{1}{\sqrt{m}}\left[\cos\left(\omega_1^\top x\right), \sin\left(\omega_1^\top x\right), \ldots\right]$.

- How fast does m need to grow with n? Sublinear for regression [Bach, 2015]

Inducing variables / Nyström

- Directly approximate the $n \times n$ Gram matrix K_{XX} of a set of inputs $\{x_i\}_{i=1}^n$ with

$$\hat{K}_{XX} = K_{XZ} K_{ZZ}^{-1} K_{ZX}$$

  where K_{ZZ} is $m \times m$ on "inducing" inputs $\{z_i\}_{i=1}^m$.
- Corresponds to explicit feature representation $x \mapsto K_{xZ} K_{ZZ}^{-1/2}$.
- Surrogate kernel $\hat{k}(x, x') = \langle k_|(\cdot, x), k_|(\cdot, x') \rangle$, where $k_|(\cdot, x)$ is a projection of $k(\cdot, x)$ to span $\{k(\cdot, z_1), \ldots, k(\cdot, z_m)\}$
- Often used in regression with Gaussian processes: with the use of Sherman-Morrison-Woodbury identity, reduces $O(n^3)$ cost to $O(nm^2)$.

  [Quiñonero-Candela and Rasmussen, 2005, Snelson and Ghahramani, 2006]
- m can grow much slower than n in regression without sacrificing performance [Rudi, Camoriano & Rosasco, 2015].

Kernel Methods – Discussion

- Kernel methods allows for very flexible and powerful machine learning models.
- **Nonparametric** method: parameter space (e.g., normal vector w in SVM) can be infinite-dimensional
- Kernels can be defined over more complex structures than vectors, e.g. graphs, strings, images, bags of instances, probability distributions.
- In naïve implementation, computational cost is at least quadratic in the number of observations, often $O(n^3)$ computation and $O(n^2)$ memory, but there are various approximations with good scaling up properties.
- Further reading:
  - Schölkopf and Smola, Learning with Kernels, 2001.
  - Rasmussen and Williams, Gaussian Processes for Machine Learning, 2006.
  - Steinwart and Christmann, Support Vector Machines, 2008.
  - Berlinet and Thomas-Agnan, Reproducing Kernel Hilbert Spaces in Probability and Statistics, 2004.
  - Bishop, Pattern Recognition and Machine Learning, Chapter 6.