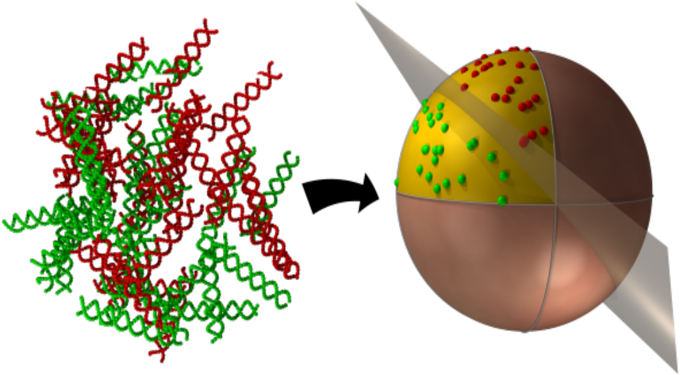
Ch7. Kernel Methods

ST4240, 2014/2015

Version 0.1

Alexandre Thiéry

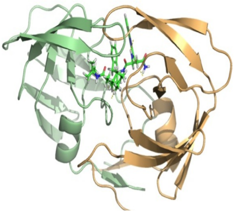
Department of Statistics and Applied Probability

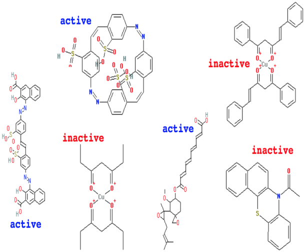


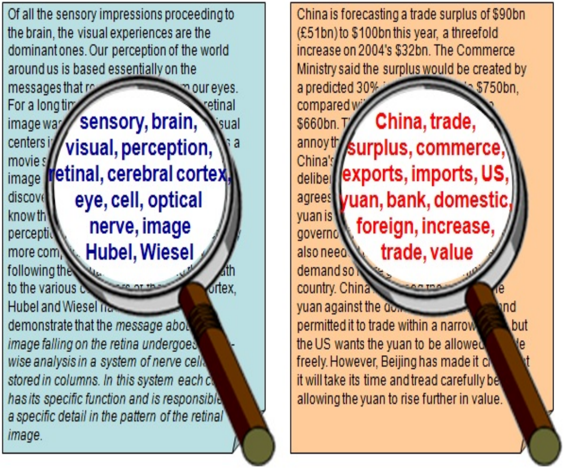
MASKATLLLAFTLLFATCIARHQQRQQQQNQCQLQNIEA...

MARSSLFTFLCLAVFINGCLSQIEQQSPWEFQGSEVW...

MALHTVLIMLSLLPMLEAQNPEHANITIGEPITNETLGWL...







Of all the sensory impressions proceeding to
the brain, the visual experiences are the
dominant ones. Our perception of the world
around us is based essentially on the
sensory, brain,
visual, perception,
retinal, cerebral cortex,
eye, cell, optical
nerve, image
Hubel, Wiesel
demonstrate that the message abo
image falling on the retina undergoes
wise analysis in a system of nerve cel
stored in columns. In this system each c
has its specific function and is responsibl
a specific detail in the pattern of the retinal
image.
China is forecasting a trade surplus of $90bn
(£51bn) to $100bn this year, a threefold
increase on 2004's $32bn. The Commerce
Ministry said the surplus would be created by
China, trade,
surplus, commerce,
exports, imports, US,
yuan, bank, domestic,
foreign, increase,
trade, value
permitted it to trade within a narrow
the US wants the yuan to be allowed
freely. However, Beijing has made it c
it will take its time and tread carefully be
allowing the yuan to rise further in value.

Outline

Regularization and Optimization

- Training examples: $(x_i, y_i)_{i=1}^N$ with $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$
- Class of functions $\mathcal{C}$, regularisation functional $\Omega(\cdot)$

$$\mathrm{argmin}_{f \in \mathcal{C}} \left\{ f \mapsto \sum_{i=1}^{N} \mathbf{Loss}\left(f(x_i), y_i\right) \; + \; \lambda \cdot \Omega(f) \right\}$$

- **[Exercise]** : OLS, Ridge regression, LASSO, Logistic regression, SVM, Boosting?

Some issues

$$\operatorname{argmin}_{f \in \mathcal{C}} \left\{ f \mapsto \sum_{i=1}^{N} \textbf{Loss}\left(f(x_i), y_i\right) \;+\; \lambda \cdot \Omega(f) \right\}$$

- Linear models: $f(x) = \langle \beta, x \rangle$ with $\beta \in \mathbb{R}^p$.
- What if the class of function $\mathcal{C}$ is very large, or infinite dimensional?
- **Example**: regression with $\mathcal{C} = (\text{Smooth functions})$ and

$$\Omega(f) = \int \left\| f' \right\|^2 (u)\, du.$$

Unstructured Data and feature extraction

- Digit classification: what features?
- Spam v.s. No Spam: what features?
- Molecule Classification: what features?

Objective

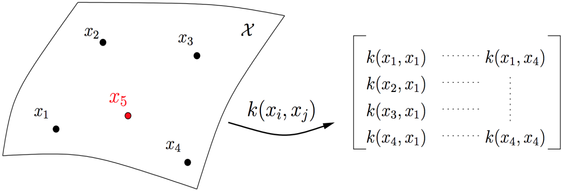
- For a general set $\mathcal{X}$, define a sensible class of functions $\mathcal{C}$
- Define the regularisation functional Ω such that the optimisation problem

$$\mathrm{argmin}_{f \in \mathcal{C}} \left\{ f \mapsto \sum_{i=1}^{N} \mathbf{Loss}\left(f(x_i), y_i\right) \; + \; \lambda \cdot \Omega(f) \right\}$$

  can be efficiently solved, even if $\mathcal{C}$ is huge.
- Understand why a large class of function $\mathcal{C}$ does not necessarily lead to bad performances (overfitting?)

Outline

Measure of similarity



- Unstructure data $\{x_i\}_{i=1}^N$
- Use a **symmetric kernel** $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to measure similarity

$$k(x_i, x_j) \approx (\text{similarity between } x_i \text{ and } x_j)$$

- Consider the **Gram matrix K** defined as

$$[K]_{i,j} = k(x_i, x_j)$$

- K contains all the pairwise measures of similarity

Positivity

- The kernel **kernel** $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semidefinite if for any $\{x_i\}_{i=1}^N$ the **Gram matrix K** defined as

$$[K]_{i,j} = k(x_i, x_j)$$

  is **positive semi-definite**.
- **[Exercise]** What about $\mathcal{X} = \mathbb{R}^d$ and $k(x_i, x_j) = \langle x_i, x_j \rangle$?
- **[Exercise]** For an arbitrary set $\mathcal{X}$ and a **feature map** $\varphi : \mathcal{X} \to \mathbb{R}^d$, what about

$$k(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle?$$

Operation on kernels

- Suppose that k_1 and k_2 are two positive semidefinite kernels on $\mathcal{X}$.
- **[Exercise]** : is $k_+(x_i, x_j) = k_1(x_i, x_j) + k_2(x_i, x_j)$ positive semidefinite?
- **[Exercise]** : is $k_\times = k_1(x_i, x_j) \times k_2(x_i, x_j)$ positive semidefinite?
- **[Exercise]** : is $k_{\exp}(x_i, x_j) = \exp\left(k(x_i, x_j)\right)$ positive semidefinite?
- **[Exercise]** : is $k_{poly}(x_i, x_j) = (1 + k(x_i, x_j))^p$ positive semidefinite?
- **Radial Basis Function kernel**: the kernel on $\mathbb{R}^d$ defined as

$$k_{RBF}(x_i, x_j) = \exp\left\{-\frac{\|x_i - x_j\|^2}{\ell^2}\right\}$$

is positive definite.

String kernel

```
IPTSALVKETLALLSTHRTLLIANETLRIPVPVHKNHQLCTEEIFQGIGTLESQTVQGGTV
ERLFKNLSLIKKYIDGQKKKCGEERRRVNQFLDYLQEFLGVMNTEWI
```

```
PHRRDLCSRSIWLARKIRSDLTALTESYVKHQGLWSELTEAERLQENLQAYRTFHVLLA
RLLEDQQVHFTPTEGDFHQAIHTLLLQVAAFAYQIEELMILLEYKIPRNEADGMLFEKK
LWGLKVLQELSQWTVRSIHDLRFISSHQTGIP
```

- One can define a kernel on strings of character. For two strings x and y define

$$\mathbf{k}(\mathbf{x}, \mathbf{y}) = \sum_{s \in \mathcal{S}} \omega_s \, \varphi_s(x) \, \varphi_s(y)$$

  where:
  - $\mathcal{S}$ a set of possible substring
  - ω_s is a positive weight
  - $\varphi_s(x)$ is the number of occurrences of s in x.
- **[Exercise]**: why is it positive semi-definite?

Outline

Reproducing Kernel Hilbert Space

$$\mathrm{argmin}_{f\in\mathcal{C}} \left\{ f \mapsto \sum_{i=1}^{N} \textbf{Loss}\left(f(x_i), y_i\right) \ + \ \lambda \cdot \Omega(f) \right\}$$

- Recall that one were interested in defining a class $\mathcal{C}$ of functions on $\mathcal{X}$ and a regularisation functional Ω
- Suppose that $k(\cdot,\cdot)$ is a positive semi-definite kernel on $\mathcal{X}$.
- For any $x_0 \in \mathcal{X}$ one can define a function $F_{x_0}(\cdot) : \mathcal{X} \to \mathbb{R}$ by

$$\mathbf{F_{x_0}(x) = k(x_0, x)}$$

Reproducing Kernel Hilbert Space

- The RKHS $\mathcal{H}$ of function is the class $\mathcal{C}$ of functions on $\mathcal{X}$ that can be expressed as

$$F(\cdot) = \sum_i \alpha_i \, F_{x_i}(\cdot)$$

for some elements $\{x_i\}_{i \in I}$ of $\mathcal{X}$

- Note that $\mathcal{H}$ vector space of functions
- One can define the norm

$$\|F\|^2_{RKHS} = \sum_{i,j} \alpha_i \, \alpha_j \, k(x_i, x_j)$$

Representer theorem

- Class of functions: $\mathcal{C} \equiv \mathcal{H}$
- Regularization: $\Omega(f) \equiv \Psi\left(\|f\|_{RKHS}\right)$ for strictly increasing Ψ
- In other words

$$\operatorname{argmin}_{f \in \mathcal{H}} \left\{ f \mapsto \sum_{i=1}^{N} \mathbf{Loss}\left(f(x_i), y_i\right) \; + \; \lambda \cdot \Psi\left(\|f\|_{\mathcal{H}}\right) \right\} \qquad (1)$$

Theorem (Representer Theorem)

A solution $f_\star$ to the optimisation problem (1) can always be expressed as

$$f_\star(\cdot) = \sum_{i=1}^{N} \alpha_i F_{x_i}(\cdot)$$

where $\{(x_i, y_i)\}_{i=1}^{N}$ are the training examples and $\{\alpha_i\}_{i=1}^{N}$ are some coefficients.

Representer theorem: consequence

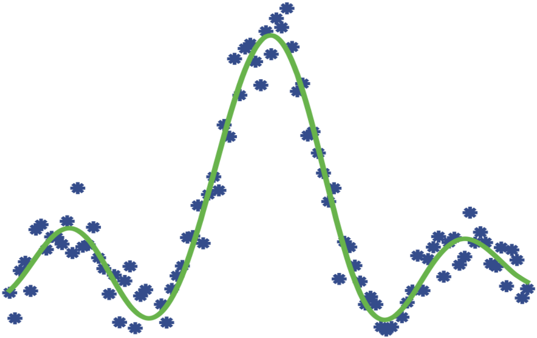
- Original optimisation problem

$$\mathrm{argmin}_{f\in\mathcal{H}} \left\{ f \mapsto \sum_{i=1}^{N} \textbf{Loss}\left(f(x_i), y_i\right) \; + \; \lambda \cdot \Psi\left(\|f\|_{\mathcal{H}}\right) \right\} \qquad (2)$$

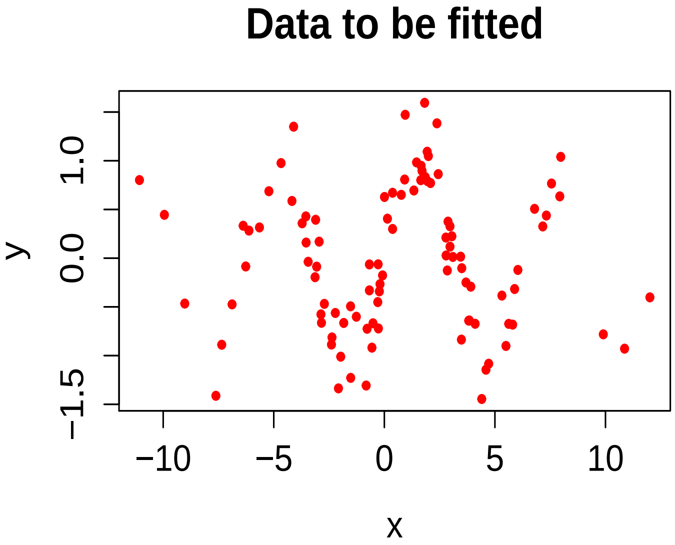
- Thanks to the representer theorem, optimisation over $f \in \mathcal{H}$ reduces to optimisation over $\alpha \in \mathbb{R}^N$

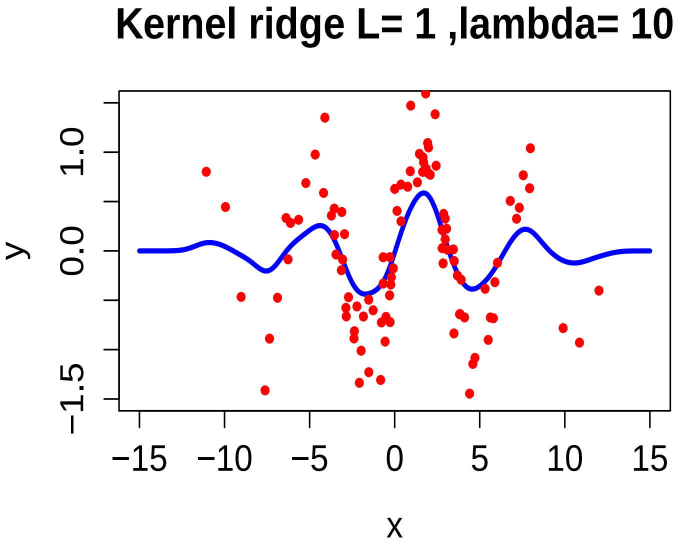
$$\mathrm{argmin}_{\alpha\in\mathbb{R}^N} \left\{ \alpha \mapsto \sum_{i=1}^{N} \textbf{Loss}\left(f(x_i), [K\alpha]_i\right) \; + \; \lambda \cdot \langle \alpha, K\alpha \rangle \right\} \qquad (3)$$

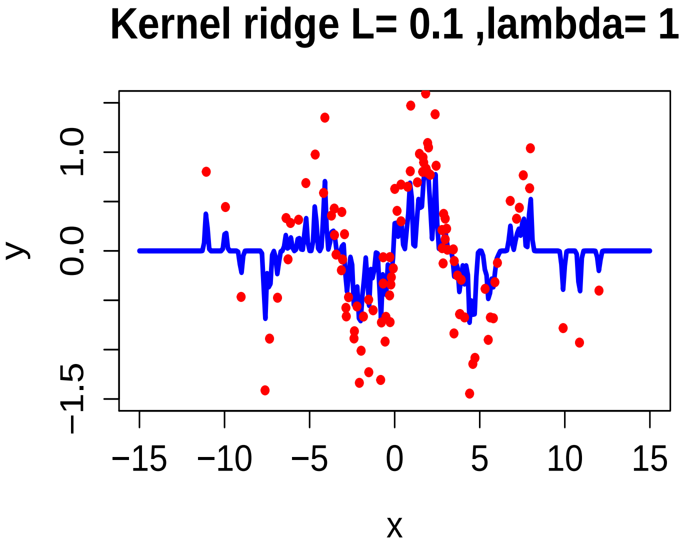
Kernel Ridge Regression

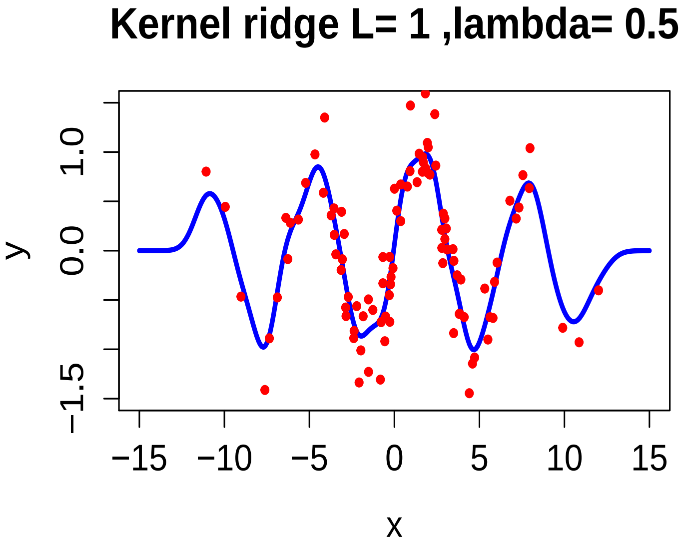


- **[Exercise]** What about ridge regression with kernel $k(\cdot, \cdot)$?

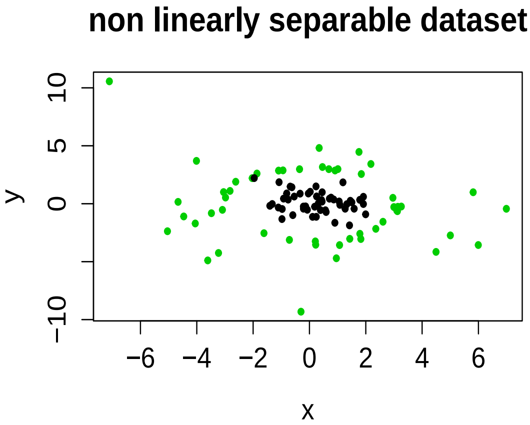

Data to be fitted
y
1.0
0.0
−1.5
−10
−5
0
5
10
x


Kernel ridge L= 1 ,lambda= 10
y
−1.5
0.0
1.0
x
−15
−10
−5
0
5
10
15


Kernel ridge L= 0.1 ,lambda= 1
y
−1.5
0.0
1.0
−15
−10
−5
0
5
10
15
x

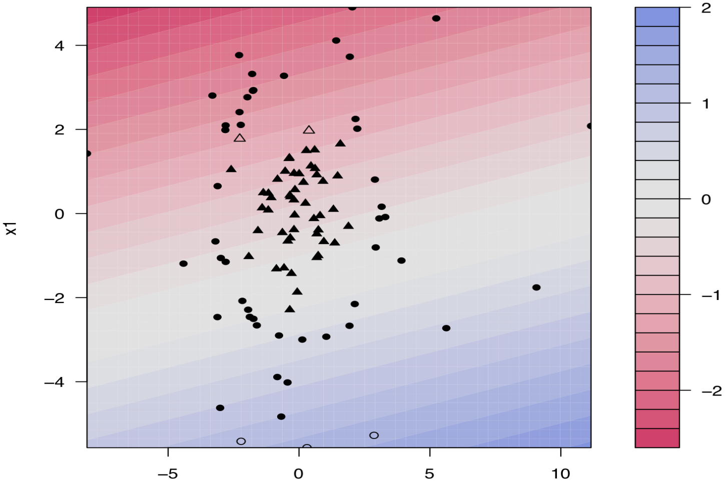

Kernel ridge L= 1 ,lambda= 0.5
y
−1.5
0.0
1.0
x
−15
−10
−5
0
5
10
15

Kernelized SVM



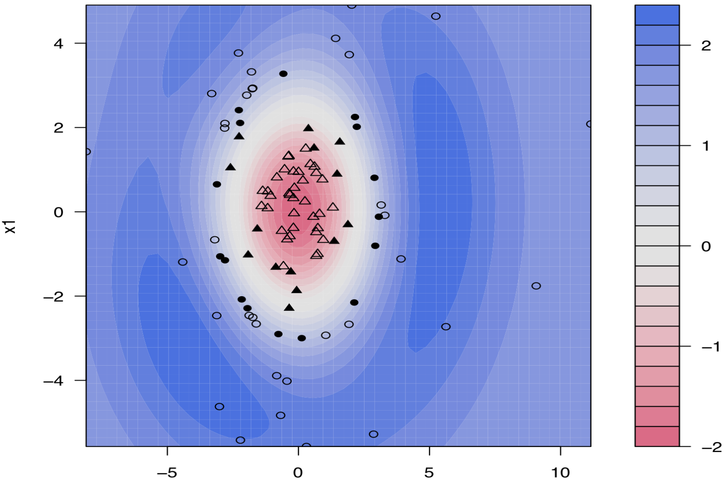
- **[Exercise]** What SVM with a kernel $k(\cdot,\cdot)$?

Kernelized SVM



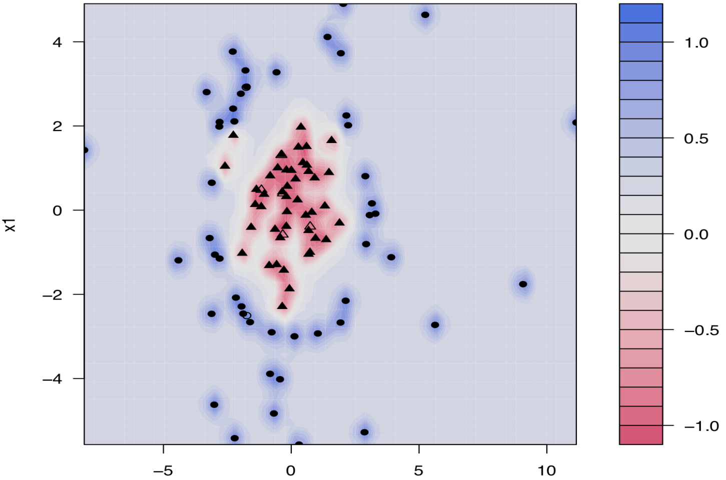
Linear Kernel

Kernelized SVM



RBF kernel with $\sigma = 1/\ell = 0.5$

Kernelized SVM



RBF kernel with $\sigma = 1/\ell = 50$