

Advanced Topics in Machine Learning (2010, Part II)

2. Regularization / Kernels Review

Massimiliano Pontil

Today's Plan

- Representer theorem
- Obtaining problems with a limited number of variables
- Kernels

Previous Lecture

We studied the problems

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^m (y_i - w^\top x_i)^2 + \gamma \|w\|^2 \quad (\text{Ridge Regression})$$

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i$$

$$\text{subject to } y_i(w^\top x_i) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \text{for } i = 1, \dots, m \quad (\text{SVM})$$

We showed that they are special cases of *regularization*

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^m E(w^\top x_i, y_i) + \gamma \|w\|^2 \quad (\mathcal{R})$$

Optimality Conditions

- For SVM, recall how the *primal* and *dual* variables are related (by KKT conditions):

$$\hat{w} = \sum_{i=1}^m \hat{\alpha}_i y_i x_i \quad (1)$$

- For ridge regression, the solution is

$$\begin{aligned} \hat{w} = (XX^\top + \gamma I_d)^{-1} Xy &\implies (XX^\top + \gamma I_d)\hat{w} = Xy \implies \\ \hat{w} = \frac{1}{\gamma} X (y - X^\top \hat{w}) &\implies \hat{w} = \sum_{i=1}^m c_i x_i \end{aligned} \quad (2)$$

where $c_i = \frac{1}{\gamma} (y_i - x_i^\top \hat{w})$

Optimality Conditions (contd.)

- In general, can we get an optimality condition like (1), (2) for the regularization problem $(\mathcal{R})$?
- Indeed, in many of these problems one may set the gradient of the Lagrangian wrt. w to zero. The resulting equation is of the form

$$\hat{w} = \sum_{i=1}^m c_i x_i$$

where $c_1, \dots, c_m \in \mathbb{R}$

- This very useful fact is known in ML as the *Representer Theorem*

Representer Theorem

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^m E(w^\top x_i, y_i) + \gamma \|w\|^2 \quad (\mathcal{R})$$

Theorem 1. *There exists a solution $\hat{w}$ of problem $(\mathcal{R})$ that has the form*

$$\hat{w} = \sum_{i=1}^m c_i x_i \quad (3)$$

for some $c_1, \dots, c_m \in \mathbb{R}$.

- In different words, there is a solution of $(\mathcal{R})$ which lies in the *linear span* of the vectors $x_1, \dots, x_m$. If $d \gg m$, this span is a subspace of $\mathbb{R}^d$ with much smaller dimensionality; thus the problem reduces to one in much fewer variables.

Representer Theorem (contd.)

Proof (in the case that E is differentiable). Let $\hat{w}$ be a solution. Setting the derivative wrt. w to zero, we have

$$\sum_{i=1}^m E'(\hat{w}^\top x_i, y_i) x_i + 2\gamma \hat{w} = 0$$

where E' is the partial derivative of E wrt. its first argument. Selecting

$$c_i = -\frac{1}{2\gamma} E'(\hat{w}^\top x_i, y_i) \tag{4}$$

we obtain (3). $\square$

- Note: this proof in fact shows that, if E is differentiable, *any* solution of $(\mathcal{R})$ is of the form (3)

Representer Theorem (contd.)

Proof (general). Let $\hat{w}$ be a solution of $(\mathcal{R})$. We can write $\hat{w}$ as

$$\hat{w} = \sum_{i=1}^m c_i x_i + n$$

where $c_1, \dots, c_m \in \mathbb{R}$ and $n \in \mathbb{R}^d$ is such that $n^\top x_i = 0$ for $i = 1, \dots, m$. In other words, we decompose $\hat{w}$ in two terms, one in the subspace spanned by the x_i and one that is orthogonal to this subspace. Then

$$\|\hat{w}\|^2 = \left\| \sum_{i=1}^m c_i x_i \right\|^2 + 2n^\top \sum_{i=1}^m c_i x_i + \|n\|^2 = \left\| \sum_{i=1}^m c_i x_i \right\|^2 + \|n\|^2 \geq \left\| \sum_{i=1}^m c_i x_i \right\|^2.$$

Also, $\hat{w}^\top x_j = (\sum_{i=1}^m c_i x_i)^\top x_j + n^\top x_j = (\sum_{i=1}^m c_i x_i)^\top x_j$ for every j ; hence the error term does not change. Thus, $\sum_{i=1}^m c_i x_i$ is a solution of $(\mathcal{R})$. $\square$

Representer Theorem (contd.)

- Plugging eq. (3) in the rhs. of eq. (4), we obtain a set of equations for the coefficients

$$c_i = -\frac{1}{2\gamma} E' \left(\sum_{j=1}^m c_j x_j^\top x_i, y_i \right) \quad i = 1, \dots, m$$

These are *necessary* conditions for the coefficients c_i to be optimal (in the case that E is differentiable).

Equivalent Problems

- The Representer Theorem provides a way to obtain a problem equivalent to a general regularization problem
 - Write a solution $\hat{w}$ in the form (3)
 - Substitute for w in $(\mathcal{R})$
- In this way we obtain an equivalent problem in m variables $c_1, \dots, c_m$
- The advantage is that if $d \gg m$ the equivalent problem can be solved in much less time than solving the original problem

Ridge Regression

- Substituting (3) in the ridge regression problem gives

$$\begin{aligned} \min_{c \in \mathbb{R}^m} \sum_{i=1}^m \left(y_i - \sum_{j=1}^m c_j x_j^\top x_i \right)^2 + \gamma \left\| \sum_{i=1}^m c_i x_i \right\|^2 &\iff \\ \min_{c \in \mathbb{R}^m} \sum_{i=1}^m \left(y_i - \sum_{j=1}^m c_j x_j^\top x_i \right)^2 + \gamma \sum_{i=1}^m c_i x_i^\top \sum_{j=1}^m c_j x_j &\quad (5) \end{aligned}$$

- Define the $m \times m$ matrix G as

$$G_{ij} = x_i^\top x_j$$

It is called the *Gram matrix* and it is *symmetric*

Ridge Regression (contd.)

- Then (5) becomes

$$\min_{c \in \mathbb{R}^m} \sum_{i=1}^m \left(y_i - \sum_{j=1}^m c_j G_{ij} \right)^2 + \gamma \sum_{i,j=1}^m c_i c_j G_{ij}$$

or

$$\min_{c \in \mathbb{R}^m} \sum_{i=1}^m (y_i - c^\top g_i)^2 + \gamma c^\top G c$$

where g_i denotes the i -th column of matrix G .

- The term $c^\top G c$ is a *quadratic form*

Ridge Regression (contd.)

- To derive the solution, we follow the same approach as for the original problem (i.e. set the derivative to zero)
- There may be infinite solutions, depending on whether the input vectors x_i are *linearly independent*, but they all generate *the same* $\hat{w}$ (we skip the proof)
- The minimal norm solution for $\hat{c}$ is

$$\hat{c} = (G + \gamma I_m)^{-1}y$$

Ridge Regression (contd.)

Algorithm

- Solve the $m \times m$ linear system

$$(G + \gamma I_m) \hat{c} = y$$

- Compute $\hat{w} = \sum_{i=1}^m \hat{c}_i x_i$

- $O(m^3 + md)$ operations $+ O(m^2d)$ to generate the Gram matrix
- Contrast with $O(d^3 + d^2m)$ operations if we solved the primal

SVM

- Substituting (3) in the SVM problem gives

$$\min_{c \in \mathbb{R}^m} \frac{1}{2} \left\| \sum_{i=1}^m c_i x_i \right\|^2 + C \sum_{i=1}^m \xi_i$$

subject to $y_i \sum_{j=1}^m c_j x_j^\top x_i \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \text{for } i = 1, \dots, m \iff$

$$\min_{c \in \mathbb{R}^m} \frac{1}{2} \sum_{i=1}^m c_i x_i^\top \sum_{j=1}^m c_j x_j + C \sum_{i=1}^m \xi_i$$

subject to $y_i \sum_{j=1}^m c_j G_{ij} \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \text{for } i = 1, \dots, m \iff$

SVM (contd.)

$$\min_{c \in \mathbb{R}^m} \frac{1}{2} c^\top G c + C \sum_{i=1}^m \xi_i$$

subject to $y_i(c^\top g_i) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \text{for } i = 1, \dots, m$

- This is a quadratic program similar to the SVM primal **(P)** from the previous lecture. Following a similar approach through the Lagrangian, we obtain the *same* dual problem **(D)**

$$\max_{\alpha \in \mathbb{R}^m} -\frac{1}{2} \alpha^\top A \alpha + \sum_{i=1}^m \alpha_i$$

subject to $0 \leq \alpha_i \leq C, \quad \text{for } i = 1, \dots, m$

where A is the $m \times m$ matrix $A = (y_i y_j G_{ij})_{i,j=1}^m$

SVM (contd.)

- The α 's are the same as in the last lecture. Recall that

$$\hat{w} = \sum_{i=1}^m \hat{\alpha}_i y_i x_i$$

- So, the optimal c_i coefficients are

$$\hat{c}_i = y_i \hat{\alpha}_i$$

- As with ridge regression, there may be infinite solutions for the c_i , if the inputs x_i are not linearly independent
- We shall describe some algorithms for SVMs in the next lecture

Regularization

- More generally, which problem do we obtain from the problem

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^m E(w^\top x_i, y_i) + \gamma \|w\|^2 \quad (\mathcal{R})$$

- Substituting the Representer Theorem (3) in $(\mathcal{R})$ yields

$$\min_{c \in \mathbb{R}^m} \sum_{i=1}^m E(c^\top g_i, y_i) + \gamma c^\top G c \quad (\mathcal{C})$$

Remarks

- Problem $(\mathcal{C})$ does not involve the inputs x_i but instead it involves the Gram matrix G
- The Gram matrix needs to be computed *once*, in $O(dm^2)$ operations
- The number of variables in problem $(\mathcal{C})$ is m ; so, if $d \gg m$ it is more efficient to solve $(\mathcal{C})$ than $(\mathcal{R})$
- The function learned is equal to

$$\hat{f}(x) = \hat{w}^\top x = \sum_{i=1}^m \hat{c}_i x_i^\top x$$

which involves only *inner products* in the input space

General Feature Map

- Alternatively, in the place of the inputs x_i we may have a linear or nonlinear *feature map* $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^N$
- $\phi(x)$ is called the *feature vector* and the space $\{\phi(x) : x \in \mathbb{R}^d\}$ the *feature space*
- That is, we fix ϕ and learn a function

$$f(x) = w^\top \phi(x)$$

- The feature space may be *high-dimensional* or even *infinite-dimensional* ($N = \infty$)

General Feature Map (contd.)

If we replace x_i with $\phi(x_i)$, we obtain the equivalent problem

$$\min_{c \in \mathbb{R}^m} \sum_{i=1}^m E(c^\top g_i, y_i) + \gamma c^\top G c \quad (\mathcal{C})$$

where

$$G_{ij} = \phi(x_i)^\top \phi(x_j) \quad \text{for all } i, j = 1, \dots, m$$

The function learned is $f(x) = \sum_{i=1}^m \hat{c}_i \phi(x_i)^\top \phi(x)$

- **Key observation:** in this representation of f , we don't need to know ϕ explicitly; we just need to know the inner product between any pair of feature vectors!

Example

- Suppose that ϕ is such that

$$\phi(x)^\top \phi(t) = e^{-\|x-t\|^2} \quad \text{for all } x, t \in \mathbb{R}^d$$

(Gaussian kernel)

- The feature map ϕ is implicit and has to be infinite-dimensional
- The function we learn, $f(x) = w^\top \phi(x)$, is nonlinear and may fit the data much better than without a feature map
- But the time cost of computing the Gram matrix and solving the equivalent problem ($\mathcal{C}$) depends only on d , which can be reasonably small

Kernels

- Let us start directly with a function K and see when K can be expressed as an inner product in some feature space, i.e.

$$K(x, t) = \langle \phi(x), \phi(t) \rangle \quad \text{for all } x, t \in \mathbb{R}^d \quad (6)$$

- **Question:** Given a function $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, which properties of K guarantee that there exists a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathbb{R}^d \rightarrow \mathcal{H}$ such that (6) holds?
- **Answer:** K should be a *positive semidefinite kernel*

Kernels

- **Definition:** A function $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a **positive semidefinite kernel** if
 - K is symmetric
 - for every $m \in \mathbb{N}$ and every $x_1, \dots, x_m \in \mathbb{R}^d$, the matrix $(K(x_i, x_j))_{i,j=1}^m$ is positive semidefinite

Regularization with Kernels

- Before discussing kernels, let us derive the equivalent regularization problem in terms of the kernel
- From $(\mathcal{C})$ and the definition $G_{ij} = \phi(x_i)^\top \phi(x_j)$, we obtain the problem

$$\min_{c \in \mathbb{R}^m} \sum_{i=1}^m E((\mathbf{K}c)_i, y_i) + \gamma c^\top \mathbf{K}c \quad (\mathcal{K})$$

where we use $\mathbf{K}$ to denote the $m \times m$ *kernel matrix* – or *Gram matrix* – on the data

$$\mathbf{K}_{ij} := K(x_i, x_j) \quad \text{for all } i, j = 1, \dots, m$$

Regularization with Kernels (contd.)

- By the Representer Theorem, the function learned equals

$$f(x) = \sum_{i=1}^m \hat{c}_i K(x_i, x)$$

- So, we only need to find the solution $\hat{c}$ of the problem $(\mathcal{K})$
- The value $f(x)$ at any point x can be computed fast as long as the kernel is easy to compute
- In the case of SVMs, the nonzero coefficients c_i correspond to the support vectors; in many situations there are few support vectors, which gives an easy to compute and interpretable solution

Linear Kernels

- The function $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ defined by

$$K(x, t) = x^\top A t \quad \text{for all } x, t \in \mathbb{R}^d,$$

where A is a $d \times d$ positive semidefinite matrix, is called a *linear kernel*

- K is a psd. kernel

Proof. Since A is psd. we can write it in the form $A = R^\top R$ for some $d \times d$ matrix R . Thus K corresponds to the feature map $\phi(x) = Rx$.

Alternatively, K is obviously symmetric and note that, for every $m \in \mathbb{N}$, $x_1, \dots, x_m \in \mathbb{R}^d$, $z \in \mathbb{R}^d$,

$$z^\top \mathbf{K} z = \sum_{i,j=1}^d z_i z_j x_i^\top A x_j = \sum_{i,j=1}^d z_i z_j (Rx_i)^\top (Rx_j) = \left\| \sum_{i=1}^d z_i Rx_i \right\|^2 \geq 0$$

Kernel Composition

- More generally, if $K : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ is a kernel and $g : \mathbb{R}^d \rightarrow \mathbb{R}^N$ is any function, then the function $\tilde{K} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$,

$$\tilde{K}(x, t) = K(g(x), g(t)) \quad \text{for all } x, t \in \mathbb{R}^d$$

is also a kernel.

Proof. Since K is symmetric, $\tilde{K}$ is clearly symmetric too. The hypothesis that K is a kernel also implies that for every $x_1, \dots, x_m \in \mathbb{R}^d$ the matrix $(K(g(x_i), g(x_j)))_{i,j=1}^m$ is psd. $\square$

- In particular, the linear kernel is a special case, corresponding to $K(x, t) = x^\top t$ and $g(x) = Rx$

Kernel Construction

- **Question:** If $K_1, \dots, K_q$ are kernels on $\mathbb{R}^d$ and $F : \mathbb{R}^q \rightarrow \mathbb{R}$, when is the function

$$F(K_1(x, t), \dots, K_q(x, t))$$

a kernel?

- We discuss some examples of functions F for which the answer to this question is YES.

Nonnegative Combination of Kernels

- If $\mu_1, \dots, \mu_q \geq 0$, then $\sum_{j=1}^q \mu_j K_j$ is a kernel.
- This fact is immediate (a nonnegative combination of psd. matrices is still psd.)

Example: Let $q = d$ and $K_i(x, t) = x_i t_i$ (linear kernel with $A = I_d$)

- In particular, this implies that
 - aK is a kernel if $a \geq 0$ and K is a kernel
 - $K_1 + K_2$ is a kernel if K_1, K_2 are kernels

Product of Kernels

- The pointwise product of two kernels K_1 and K_2

$$K(x, t) := K_1(x, t)K_2(x, t)$$

is a kernel.

Proof. We need to show that if A and B are $m \times m$ psd. matrices, so is $C = (A_{ij}B_{ij})_{i,j=1}^m$ (C is also called the *Schur product* of A and B). We write A and B in their singular value decomposition $A = U\Sigma U^\top$, $B = V\Lambda V^\top$ where U, V are orthogonal matrices and $\Sigma = \text{Diag}(\sigma_1, \dots, \sigma_r)$, $\Lambda = \text{Diag}(\lambda_1, \dots, \lambda_s)$, $r, s \leq m$. We have

$$\sum_{i,j=1}^m z_i z_j C_{ij} = \sum_{i,j=1}^m z_i z_j \sum_{k=1}^r \sigma_k U_{ik} U_{jk} \sum_{\ell=1}^s \lambda_\ell V_{i\ell} V_{j\ell} =$$

Product of Kernels (contd.)

$$\sum_{k=1}^r \sum_{\ell=1}^s \sigma_k \lambda_\ell \sum_{i=1}^m z_i U_{ik} V_{i\ell} \sum_{j=1}^m z_j U_{jk} V_{j\ell} = \sum_{k=1}^r \sum_{\ell=1}^s \sigma_k \lambda_\ell \left(\sum_{i=1}^m z_i U_{ik} V_{i\ell} \right)^2 \geq 0$$

□

Summary of Kernel Properties

The above results can be summarized as follows:

If K_1, K_2 are kernels on $\mathbb{R}^d$, $a \geq 0$, K a kernel on $\mathbb{R}^N$ and $g : \mathbb{R}^d \rightarrow \mathbb{R}^N$ then the following functions are kernels on $\mathbb{R}^d$

1. $K_1(x, t) + K_2(x, t)$

2. $aK_1(x, t)$

3. $K_1(x, t)K_2(x, t)$

4. $K(g(x), g(t))$

Polynomial Kernels

- Let $F = p$ where $p : \mathbb{R}^q \rightarrow \mathbb{R}$ is a *polynomial* with *nonnegative coefficients*. By properties 1, 2 and 3 above we conclude that p is a valid function.

- In particular, for $q = 1$,

$$\sum_{i=1}^d a_i (K(x, t))^i$$

is a kernel if $a_1, \dots, a_d \geq 0$

Polynomial Kernels (contd.)

- E.g., if $a \geq 0$, the following are valid polynomial kernels

1. $(x^\top t)^s$

2. $(a + x^\top t)^s$

3. $\sum_{i=0}^s \frac{a^i}{i!} (x^\top t)^i$

'Infinite Polynomial' Kernel

- In the last equation, as $s \rightarrow \infty$ the series $\sum_{i=0}^s \frac{a^i}{i!} (x^\top t)^i$ converges everywhere uniformly to

$$e^{a x^\top t}$$

showing that this function is also a kernel (*dot product* kernel)

- Assume for simplicity that $d = 1$. A feature map corresponding to the kernel e^{axt} is

$$\phi(x) = \left(1, \sqrt{a}x, \sqrt{\frac{a}{2}}x^2, \sqrt{\frac{a^3}{6}}x^3, \dots \right) = \left(\sqrt{\frac{a^i}{i!}}x^i \right)_{i \in \mathbb{N}}$$

- The feature space has infinite dimensionality!

Translation Invariant and Radial Kernels

We say that a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is

- *Translation invariant* if it has the form

$$K(x, t) = H(x - t)$$

where $H : \mathbb{R}^d \rightarrow \mathbb{R}$ is a differentiable function.

- *Radial* if it has the form

$$K(x, t) = h(\|x - t\|)$$

where $h : [0, +\infty) \rightarrow [0, +\infty)$ is a differentiable function

- **Note:** Clearly, additional conditions on H and h should be satisfied

The Gaussian Kernel

- An important example of a radial kernel is the *Gaussian kernel*

$$K(x, t) = e^{-\beta \|x-t\|^2}$$

where $\beta > 0$

- It is a kernel because it is the product of two kernels

$$K(x, t) = e^{-\beta(x^\top x + t^\top t)} e^{2\beta x^\top t}$$

(We saw before that $\exp(2\beta x^\top t)$ is a kernel. Clearly $\exp(-\beta(x^\top x + t^\top t))$ is a kernel with one dimensional feature map $\phi(x) = \exp(-\beta x^\top x)$)

- **Exercise:** Can you find a feature map representation for the Gaussian kernel in the case $d = 1$?

Bibliography

Lectures available at:

<http://www.cs.ucl.ac.uk/staff/a.argyriou/courses/index.html>

See also Chapters 2 and 3 of Shawe-Taylor and Cristianini, Kernel Methods for Pattern Analysis, 2004.