Lecture 09-b: Support Vector Regression in some details

Instructor: Prof. Ganesh Ramakrishnan

KKT and Dual for SVR

- $\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_i(\xi_i + \xi_i^*)$
  s.t. $\forall i$,
  $y_i - w^\top \phi(x_i) - b \leq \epsilon + \xi_i$,
  $b + w^\top \phi(x_i) - y_i \leq \epsilon + \xi_i^*$,
  $\xi_i, \xi_i^* \geq 0$
- Let's consider the lagrange multipliers α_i, α_i^*, μ_i and μ_i^* corresponding to the above-mentioned constraints respectively.

KKT conditions

- Differentiating the Lagrangian w.r.t. w,
  $w - \alpha_i \phi(x_i) + \alpha_i^* \phi(x_i) = 0$
  i.e. $w = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \phi(x_i)$
- Differentiating the Lagrangian w.r.t. ξ_i,
  $C - \alpha_i - \mu_i = 0$
  i.e. $\alpha_i + \mu_i = C$
- Differentiating the Lagrangian w.r.t ξ_i^*,
  $\alpha_i^* + \mu_i^* = C$
- Differentiating the Lagrangian w.r.t b,
  $\sum_i (\alpha_i^* - \alpha_i) = 0$
- Complimentary slackness:
  $\alpha_i (y_i - w^\top \phi(x_i) - b - \epsilon - \xi_i) = 0$
  $\mu_i \xi_i = 0$
  $\alpha_i^* (b + w^\top \phi(x_i) - y_i - \epsilon - \xi_i^*) = 0$
  $\mu_i^* \xi_i^* = 0$

Conclusions from the KKT conditions:

$$\alpha_i \in (0, C) \Rightarrow ?$$

$$\alpha_i^* \in (0, C) \Rightarrow ?$$

- The primal objective and constraints are convex $\Rightarrow$ KKT conditions here necessary and sufficient and strong duality holds
- $w = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\phi(x_i) \Rightarrow$ the final decision function $f(x) = w^T\phi(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\phi^T(x_i)\phi(x)$
- The dual optimization problem to compute the α's for SVR is:

$$max_{\alpha_i,\alpha_i^*} - \frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(x_i)\phi(x_j)$$

$$-\epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$$

s.t.

  - $\sum_i(\alpha_i - \alpha_i^*) = 0$
  - $\alpha_i, \alpha_i^* \in [0, C]$

- ***We notice that the only way these three expressions involve ϕ is through $\phi^\top(x_i)\phi(x_j) = K(x_i, x_j)$, for some i, j***

How about Ridge Regression?

- Recall for Ridge Regression: $w = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T y$, where,

$$\Phi = \begin{bmatrix} \phi_1(x_1) & \dots & \phi_p(x_1) \\ \dots & \dots & \dots \\ \phi_1(x_m) & \dots & \phi_p(x_m) \end{bmatrix}$$

  and

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \dots \\ y_m \end{bmatrix}$$

- $\left(\Phi^T\Phi\right)_{ij} = \sum_{k=1}^{m} \phi_i(x_k)\phi_j(x_k)$ whereas $\left(\Phi\Phi^T\right)_{ij} = \sum_{k=1}^{p} \phi_k(x_i)\phi_k(x_j) = K(x_i, x_j)$

How about Ridge Regression?

- Given $w = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T y$ and using the identity $(P^{-1} + B^T R^{-1} B)^{-1} B^T R = PB^T(BPB^T + R)^{-1}$
  - $\Rightarrow w = \Phi^T(\Phi\Phi^T + \lambda I)^{-1} y = \sum_{i=1}^{m} \alpha_i \phi(x_i)$ where $\alpha_i = \left((\Phi\Phi^T + \lambda I)^{-1} y\right)_i$
  - $\Rightarrow$ the final decision function $f(x) = \phi^T(x)w = \sum_{i=1}^{m} \alpha_i \phi^T(x)\phi(x_i)$
- Again, ***We notice that the only way the decision function*** $f(x)$ ***involves*** ϕ ***is through*** $\phi^\top(x_i)\phi(x_j)$***, for some*** i, j

The Kernel function in Ridge Regression

- We call $\phi^\top(x_1)\phi(x_2)$ a ***kernel* function**:
  $K(x_1, x_2) = \phi^\top(x_1)\phi(x_2)$
- The preceding expression for decision function becomes
  $f(x) = \sum_{i=1}^{m} \alpha_i K(x, x_i)$
  where $\alpha_i = \left(([K(x_i, x_j)] + \lambda I)^{-1} y\right)_i$

The Kernel function in SVR

- Again, involving the ***kernel* function**: $K(x_1, x_2) = \phi^\top(x_1)\phi(x_2)$
- The dual problem becomes:

$$max_{\alpha_i,\alpha_i^*} -\frac{1}{2}\sum_i\sum_j(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i,x_j)$$

$$-\epsilon\sum_i(\alpha_i+\alpha_i^*)+\sum_i y_i(\alpha_i-\alpha_i^*)$$

  s.t.
  - $\sum_i(\alpha_i-\alpha_i^*)=0$
  - $\alpha_i, \alpha_i^* \in [0, C]$
- The decision function becomes: $f(x) = \sum_i(\alpha_i-\alpha_i^*)K(x_i,x)+b$
- We will see that, often, computing $K(x_1, x_2)$ does not even require computing $\phi(x_1)$ or $\phi(x_2)$ explicitly

An example

- Let $K(x_1, x_2) = (1 + x_1^\top x_2)^2$
- What $\phi(x)$ will give $\phi^\top(x_1)\phi(x_2) = K(x_1, x_2) = (1 + x_1^\top x_2)^2$
- Is such a ϕ guaranteed to exist?
- Is there a unique ϕ for given K?

- We can prove that such a ϕ exists
- For example, for a 2-dimensional x_i:

$$\phi(x_i) = \begin{bmatrix} 1 \\ x_{i1}\sqrt{2} \\ x_{i2}\sqrt{2} \\ x_{i1}x_{i2}\sqrt{2} \\ x_{i1}^2 \\ x_{i2}^2 \end{bmatrix}$$

- $\phi(x_i)$ exists in a 5-dimensional space
- Thus, to compute $K(x_1, x_2)$, all we need is $x_1^\top x_2$, and there is no need to compute $\phi(x_i)$

Introduction to the Kernel Trick (more later)

- **Kernels** operate in a *high-dimensional*, *implicit* feature space without ever computing the coordinates of the data in that space, but rather by simply computing the Kernel function
- This approach is called the "*kernel trick*" and will talk about *valid kernels* in the next class
- This operation is often computationally cheaper than the explicit computation of the coordinates

Sequential Minimal Optimization (SMO) for SVR

- It can be shown that the objective:
  $max_{\alpha_i,\alpha_i^*} - \frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(x_i)\phi(x_j)$
  $-\epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$
- can be written as:
  $max_{\beta_i} - \frac{1}{2}\sum_i \sum_j \beta_i \beta_j \phi^\top(x_i)\phi(x_j) - \epsilon \sum_i |\beta_i| + \sum_i y_i \beta_i$
  s.t.
  - $\sum_i \beta_i = 0$
  - $\beta_i \in [-C, C], \forall i$
- The SMO subroutine can be defined as:
  1. Initialise $\beta_1, \ldots, \beta_n$ to some value $\in [-C, C]$
  2. Pick β_i, β_j to estimate next (i.e. estimate β_i^{new}, β_j^{new})
  3. Check if the KKT conditions are satisfied
     - ★ If not, choose β_i and β_j that worst violate the KKT conditions and reiterate

Least Squares SVM

- LS-SVM gives an SVR formulation that gives closed form solution just like linear or ridge regression (since SVR deals with a continous valued predicition)
- $\min_{w,b} \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{n}(y_i - (w^\top \phi(x_i) + b))^2$
- Here, $\epsilon = 0$
- Its difference with Ridge regression is that here b is not captured within w, and b is not minimized as $\|w\|^2$ is

Solution of LS-SVM

- The objective function is convex in w and b
- Thus, $\nabla_{w,b} L(w^*, b^*) = 0$ is a necessary and sufficient condition for optimality
- w.r.t w, we have:
  $w + 2\sum_i \sum_j (\phi^\top(x_i)\phi(x_j))w + 2\sum_i (y_i - b)\phi(x_i) = 0$
- w.r.t b, we have:
  $nb + \sum_i (\phi^\top(x_i)w - y_i) = 0$
- Unlike previous formulations which had linear inequalities here we have only linear equalities, which can be solved

Thus, we obtain the closed form solution:

$$w = (K^\top K + \frac{1}{C} \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix})^{-1} \phi^\top y$$

where

- $\phi_i = \phi(x_i)$
- $K_{ij} = \phi^\top(x_i)\phi(x_j) = K(x_i, x_j)$

- LS-SVM gives us a closed-form expression for *w*. But this *"speed"* is only possible for linear kernels (which have ϕ computed anyway). No implicit computation of K for a higher dimensional ϕ is possible
- We will make a similar observation for SVM for classification, where a linear time algorithm can be formulated for linear SVM

For a given K, how to show that ϕ exists, without constructing a ϕ?

- *Mercer kernel*
- *Positive-definite kernel*
- The *Mercer kernel* and *Positive-definite kernel* turn out to be equivalent definitions of kernel if the input space $\{x\}$ is *compact* (every Cauchy sequence is convergent).

Mercer's theorem

- **Mercer kernel:** $K(x_1, x_2)$ is a Mercer kernel if $\int\int K(x_1, x_2)g(x_1)g(x_2)\, dx_1 dx_2 \geq 0$ for all square integrable functions $g(x)$
(that is, $\int (g(x))^2\, dx$ is finite)
- **Mercer's theorem:**
An implication of the theorem is that
for any *Mercer kernel* $K(x_1, x_2)$, $\exists\, \phi(x) : \mathbb{R}^n \mapsto H$,
s.t. $K(x_1, x_2) = \phi^\top(x_1)\phi(x_2)$
where H is a *Hilbert space*, which is an inner product space with associated norms, where every Cauchy sequence is convergent

Do you know Hilbert? No? Then what are you doing in his space? :)

Prove that $(x_1^\top x_2)^d$ is a Mercer kernel ($d \in \mathbb{Z}^+$, $d \geq 1$)

- We want to prove that
  $\int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2)\, dx_1 dx_2 \geq 0$,
  for all square integrable functions $g(x)$
- Here, x_1 and x_2 are vectors
- Thus, $\int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2)\, dx_1 dx_2$

$$= \int_{x_{11}} .. \int_{x_{1t}} \int_{x_{21}} .. \int_{x_{2t}} [\sum_{n_1..n_t} \frac{d!}{n_1!..n_t!} \prod_{j=1}^{t} (x_{1j} x_{2j})^{n_j}]\, g(x_1) g(x_2)\, dx_{11}..dx_{1t} dx_{21}..dx_{2t}$$

s.t. $\sum_i n_i = d$

(taking a leap)

Prove that $(x_1^\top x_2)^d$ is a Mercer kernel ($d \in \mathbb{Z}^+$, $d \geq 1$)

$$= \sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \int_{x_1} \int_{x_2} \prod_{j=1}^{t} (x_{1j} x_{2j})^{n_j} \, g(x_1) g(x_2) \, dx_1 \, dx_2$$

$$= \sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \int_{x_1} \int_{x_2} (x_{11}^{n_1} x_{12}^{n_2} \ldots x_{1t}^{n_t}) g(x_1) \, (x_{21}^{n_1} x_{22}^{n_2} \ldots x_{2t}^{n_t}) g(x_2) \, dx_1 \, dx_2$$

$$= \sum_{n_1 \ldots n_t} \frac{d!}{n_1! \ldots n_t!} \left(\int_{x_1} (x_{11}^{n_1} \ldots x_{1t}^{n_t}) g(x_1) \, dx_1 \right) \left(\int_{x_2} (x_{21}^{n_1} \ldots x_{2t}^{n_t}) g(x_2) \, dx_2 \right)$$

(integral of decomposable product as product of integrals)

s.t. $\sum_i n_i = d$

Prove that $(x_1^\top x_2)^d$ is a Mercer kernel ($d \in \mathbb{Z}^+$, $d \geq 1$)

- Realize that both the integrals are basically the same, with different variable names
- Thus, the equation becomes:

$$\sum_{n_1 \dots n_t} \frac{d!}{n_1! \dots n_t!} \left(\int_{x_1} (x_{11}^{n_1} \dots x_{1t}^{n_t}) g(x_1) \, dx_1 \right)^2 \geq 0$$

(the square is non-negative for reals)

- Thus, we have shown that $(x_1^\top x_2)^d$ is a Mercer kernel.

What about $\sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d$ s.t. $\alpha_d \geq 0$?

- $K(x_1, x_2) = \sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d$
- Is $\int_{x_1} \int_{x_2} (\sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d) \, g(x_1) g(x_2) \, dx_1 dx_2 \geq 0$?
- We have

$$\int_{x_1} \int_{x_2} (\sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d) \, g(x_1) g(x_2) \, dx_1 dx_2$$

$$= \sum_{d=1}^{r} \alpha_d \int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2) \, dx_1 dx_2$$

What about $\sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d$ s.t. $\alpha_d \geq 0$?

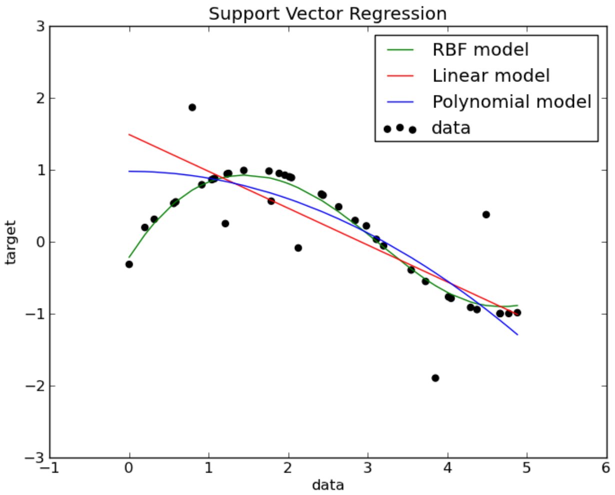
- We have already proved that $\int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2)\, dx_1 dx_2 \geq 0$
- Also, $\alpha_d \geq 0$, $\forall d$
- Thus,

$$\sum_{d=1}^{r} \alpha_d \int_{x_1} \int_{x_2} (x_1^\top x_2)^d g(x_1) g(x_2)\, dx_1 dx_2 \geq 0$$

- By which, $K(x_1, x_2) = \sum_{d=1}^{r} \alpha_d (x_1^\top x_2)^d$ is a Mercer kernel.

Kernels in SVR

- Note that the dual:
  $max_{\alpha_i,\alpha_i^*} - \frac{1}{2}\sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\phi^\top(x_i)\phi(x_j) - \epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*)$
  and the decision function:
  $f(x) = \sum_i (\alpha_i - \alpha_i^*)\phi^\top(x_i)\phi(x) + b$
  are all in terms of the dot product $\phi^\top(x_i)\phi(x_j)$ only
- *Therefore, one could employ kernels in SVR to implicitly perform linear regression in higher dimensional spaces*


Support Vector Regression
RBF model
Linear model
Polynomial model
data
target
data
3
2
1
0
−1
−2
−3
−1
0
1
2
3
4
5
6