Recitation 5

Kernels

Colin

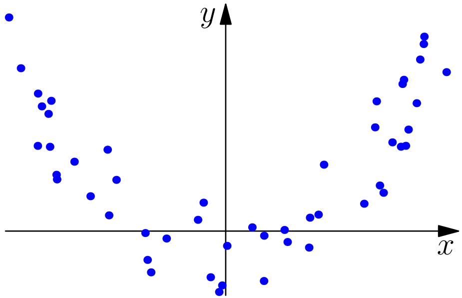
Spring 2022

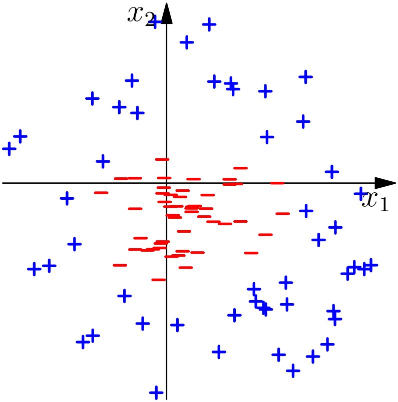
Feb 23

Motivation

Question

Consider applying linear regression to the data set on the left, and an SVM to the data set on the right. What is the issue?

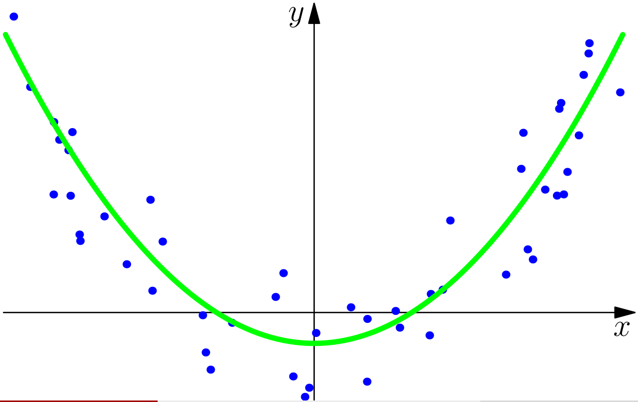




Motivation

Regression Solution

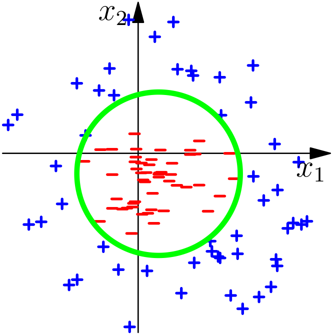
Using features $(1, x, x^2)$ and $w = (-.1, 0, 1)$ gives us $f_w(x) = -.1 + 0x + 1x^2 = x^2 - .1$. Our prediction function is quadratic but we obtained it through standard linear methods.



Motivation

SVM Solution

For the SVM we expand our feature vector from $(1, x_1, x_2)$ to $(1, x_1, x_2, x_1x_2, x_1^2, x_2^2)$. Using $w = (-1.875, 2.5, -2.5, 0, 1, 1)$ gives $-1.875 + 2.5x_1 - 2.5x_2 + x_1^2 + x_2^2 = (x_1 + 1.25)^2 + (x_2 - 1.25)^2 - 5 = 0$ as our decision boundary.



Motivation

- Linear model is clearly insufficient to represent these problems.
- The most intuitive solution is to **expand the input space**
    - Adding features
- We can define a **feature map function** $\varphi(x) : \mathcal{X} \mapsto \mathcal{H}$
    - $dim(\mathcal{H}) > dim(\mathcal{X})$
    - For ridge regression, $\varphi(1, x) = [1, x, x^2]$.
    - For SVM, $\varphi(1, x_1, x_2) = [1, x_1, x_2, x_1 x_2, x_1^2, x_2^2]$.
- We then find a linear separator on the feature space $\mathcal{H}$.

Adding Features

- From undergrad Calc (Taylor's Thm), we learned polynomials can approximate any function.
- We can linearly model any problem perfectly if we add enough terms.
- But adding features obviously comes with a cost.
- The cost grows exponentially as we increase the degree.

Adding Features

Question

Suppose we begin with d-dimensional inputs $x = (x_1, \ldots, x_d)$. We add all features up to degree M. More precisely, all terms of the form

$$x_1^{p_1} \cdots x_d^{p_d} \quad p_i \geq 0 \text{ and } p_1 + \cdots + p_d \leq M$$

How many features will we have in total?

- There will be $\binom{M+d}{M}$ terms total. If M is fixed and we let d grow, this behaves like $\frac{d^M}{M!}$
- Both M and d impacts the cost of adding features.
- If we stick with polynomial features up to order M, it's takes exponential time $O(d^M)$ to compute all features.
- **What if we don't want to reduce the model complexity? How do we make the computation feasible?**

Representer Theorem (Baby Version)

Theorem ((Baby) Representer Theorem)

Suppose you have a loss function of the form

$$J(w) = L(w^T\varphi(x_1), \dots, w^T\varphi(x_n)) + R(\|w\|_2)$$

where

- $x_i \in \mathbb{R}^d, w \in \mathbb{R}^{d'}, \varphi(x) : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$.
- $L : \mathbb{R}^n \to \mathbb{R}$ *is an arbitrary function (loss term).*
- $R : \mathbb{R}_{\geq 0} \to \mathbb{R}$ *is increasing (regularization term).*

Assume J has at least one minimizer. Then J has a minimizer w^* of the form $w^* = \sum_{i=1}^n \alpha_i \varphi(x_i)$ for some $\alpha \in \mathbb{R}^n$. If R is strictly increasing, then all minimizers have this form.

Representer Theorem

Representer Theorem: Proof

Proof.

- Let $w^* \in \mathbb{R}^{d'}$ and let $S = \text{Span}(\varphi(x_1), \ldots, \varphi(x_n))$.
- Suppose w^* is the optimal parameter, and it **does not lie in** S.
- Then we can write $w^* = u + v$ where $u \in S$ and $v \in S^\perp$. (Here u is the orthogonal projection of w^* onto S, and $S^\perp$ is the subspace of all vectors orthogonal to S.)
- Then $(w^*)^T\varphi(x_i) = (u+v)^T\varphi(x_i) = u^T\varphi(x_i) + v^T\varphi(x_i) = u^T\varphi(x_i)$. So the prediction only depends on $u^T\varphi(x_i)$.
- But $\|w^*\|_2^2 = \|u+v\|_2^2 = \|u\|_2^2 + \|v\|_2^2 + 2u^Tv = \|u\|_2^2 + \|v\|_2^2 \geq \|u\|_2^2$.
- Thus $R(\|w^*\|_2) \geq R(\|u\|_2)$ showing $J(w^*) \geq J(u)$.

□

Representer Theorem

- If your loss function only depends on w via its inner products with the inputs, and the regularization is an increasing function of the ℓ_2 norm, then we can write w^* as a linear combination of the training data.

The Kernel Function

Definition (Kernel)

Given a feautre map $\varphi(x) : \mathcal{X} \mapsto \mathcal{Z}$, the **kernel function** corresponding to $\varphi(x)$ is

$$k(x, x^{'}) = \langle \varphi(x), \varphi(x^{'}) \rangle$$

where $\langle \cdot, \cdot \rangle$ is an inner product operator.

- So a kernel function computes the inner product of applying the feature map $\varphi(x)$ for two inputs $x, x' \in \mathcal{X}$.
- We only need to know the output of the kernel to find the parameters.
- Predictor function is:

$$f(x^*) = \sum_i \alpha_i k(x_i, x^*)$$

Efficiency of Kernel

Consider the polynomial kernel $k(x, y) = \langle \varphi(x), \varphi(y) \rangle = (1 + x^T y)^M$ where $x, y \in \mathbb{R}^d$. For example, if $M = 2$ we have

$$\begin{aligned} (1 + x^T y)^2 &= 1 + 2x^T y + x^T y x^T y \\ &= 1 + 2\sum_{i=1}^d x_i y_i + \sum_{i,j=1}^d x_i y_i x_j y_j. \end{aligned}$$

Option 1: First explicitly evaluate $\varphi(x)$ and $\varphi(y)$, and then compute $\langle \varphi(x), \varphi(y) \rangle$.

- $\varphi(x) = (1, \sqrt{2}x_1, \ldots, \sqrt{2}x_d, x_1^2, \ldots, x_d^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \ldots, \sqrt{2}x_{d-1}x_d)$
- Takes $O(d^M)$ times to evaluate $\varphi(x)$ and $\varphi(y)$.
- Takes another $O(d^M)$ times to compute the inner product.
- Time complexity is $O(d^M)$.

Efficiency of Kernel

Consider the polynomial kernel $k(x, y) = \langle \varphi(x), \varphi(y) \rangle = (1 + x^T y)^M$ where $x, y \in \mathbb{R}^d$. This computes the inner product of all monomials up to degree M in time $O(d)$. For example, if $M = 2$ we have

$$\begin{aligned}(1 + x^T y)^2 &= 1 + 2x^T y + x^T y x^T y \\ &= 1 + 2\sum_{i=1}^{d} x_i y_i + \sum_{i,j=1}^{d} x_i y_i x_j y_j.\end{aligned}$$

Option 2: First calculate $1 + x^T y$, then calculate $(1 + x^T y)^M$.

- Takes $O(d)$ time to evaluate $1 + x^T y$.
- Takes $O(1)$ time to calculate $(1 + x^T y)^M$
- Time complexity is $O(d)$

Recap on what we achieved

- Start with a low dimensional model
  - Due to limited input data size
  - Number of parameters is d
- Want to increase the model capacity by adding features $x_i \to \varphi(x_i)$
  - The cost is too high as we increase degrees
  - Number of parameters is $d', d' >> d$
- Realize the optimal parameter is a linear combination of $\varphi(x_i)$
  - Representer Theorem
  - Number of parameters becomes $N, d' >> N > d$
- Realize we only need the inner produce of two $\varphi(x_i)$, $k(\cdot, \cdot)$
  - We don't need to compute $\varphi(\cdot)$
  - Greatly reduces computation cost
- The rephrased problem becomes a linear problem
  - But the solution still has high dimensional expressive power!

Mercer's Theorem

- Not all function $f(x, y)$ are valid kernels. Why?
- $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$
- How can we know if $k(x, y)$ is a valid kernel or not?

Theorem (Mercer's Theorem)

Fix a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. There is a Hilbert space H and a feature map $\varphi : \mathcal{X} \to H$ such that $k(x, y) = \langle \varphi(x), \varphi(y) \rangle_H$ if and only if for any $x_1, \ldots, x_n \in \mathcal{X}$ the associated matrix K is positive semi-definite:

$$K = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}.$$

Such a kernel k is called **positive semi-definite**.

Positive Semi-Definite

Definition (Positive Semi-Definite)

A matrix $A \in \mathbb{R}^{n \times n}$ is **positive semi-definite** if it is symmetric and

$$x^T A x \geq 0$$

for all $x \in \mathbb{R}^n$.

- Equivalent to saying the matrix is symmetric with non-negative eigenvalues.

Valid Kernels

A function $k(x, y)$ is a valid kernel iff it satisfies all the properties of inner product:

- Symmetricity
  - $k(x, y) = k(y, x)$.
- Non-negativity
  - $k(x, x) \geq 0$, equality holds when $x = 0$
- Linearity
  - $k(ax, by) = abk(x, y)$
- OR The Gram Matrix K is positive semi-definitive.

Kernel Examples

- Dot Product
  - $k(x_i, x_j) = x_i^T x_j$
- Mth Polynomial Kernels
  - $k(x_i, x_j) = (1 + x_i^T x_j)^M$
- RBF Kernels
  - $k(x_i, x_j) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2})$
- Sigmoid kernel
  - $k(x_i, x_j) = tanh(\alpha x_i^T x_j + c)$

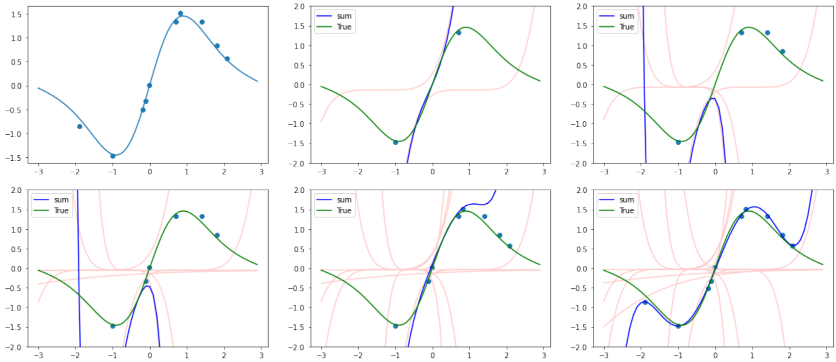
Going to infinite dimension

- What is the polynomial expression of $\varphi(\cdot)$ for RBF and Sigmoid Kernel?
  - There are no finite expression, they are sum of infinite polynomials
  - $\varphi(x) = e^{-x^2/2\sigma^2}\left[1, \sqrt{\frac{1}{1!\sigma^2}}x, \sqrt{\frac{1}{2!\sigma^4}}x^2, \sqrt{\frac{1}{3!\sigma^6}}x^3, \ldots\right]$
- This implies we have essentially modeled the problem using a infinite degree polynomial!
- At this point, the factor limiting our model capacity is the amount of training data.

What are Kernels doing

What are Kernels doing

$$f(x) = sin(x)e^{cos(x)}$$



Representer Theorem: Ridge Regression

- By adding features to ridge regression we had

$$\begin{aligned} J(\tilde{w}) &= \frac{1}{n}\sum_{i=1}^{n}(\tilde{w}^T\varphi(x_i) - y_i)^2 + \lambda\|\tilde{w}\|_2^2 \\ &= \frac{1}{n}\|\Phi\tilde{w} - y\|_2^2 + \lambda\tilde{w}^T\tilde{w}, \end{aligned}$$

  where $\Phi \in \mathbb{R}^{n\times d'}$ is the matrix with $\varphi(x_i)^T$ as its ith row.
- Representer Theorem applies giving $\tilde{w} = \sum_{j=1}^{n}\alpha_j\varphi(x_j) = \Phi^T\alpha$.
- Plugging in gives

$$J(\alpha) = \frac{1}{n}\left\|\Phi\Phi^T\alpha - y\right\|_2^2 + \lambda\alpha^T\Phi\Phi^T\alpha.$$

- Define $K = \Phi\Phi^T$

Representer Theorem: Primal SVM

- For a general linear model, the same derivation above shows

$$J(w) = L(\Phi w) + R(\|w\|_2)$$

  becomes

$$J(\alpha) = L(K\alpha) + R(\sqrt{\alpha^T K\alpha}).$$

  Here $\varphi(x_i)^T w$ became $(K\alpha)_i$.

- The primal SVM has loss function

$$J(w) = \frac{c}{n}\sum_{i=1}^{n}(1 - y_i(\varphi(x_i)^T w))_+ + \|w\|_2^2.$$

- This is kernelized to

$$J(\alpha) = \frac{c}{n}\sum_{i=1}^{n}(1 - y_i(K\alpha)_i)_+ + \alpha^T K\alpha.$$

- Positive decision made if $(w^*)^T\varphi(x) = \sum_{i=1}^n \alpha_i k(x_i, x) > 0$.

Dual SVM

- The dual SVM problem (with features) is given by

$$
\begin{aligned}
\text{maximize}_\alpha \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \varphi(x_i)^T \varphi(x_j) \\
\text{subject to} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \alpha_i \in \left[0, \frac{c}{n}\right] \quad \text{for } i = 1, \ldots, n.
\end{aligned}
$$

- We can immediately kernelize (no representer theorem needed) by replacing $\varphi(x_i)^T \varphi(x_j) = k(x_i, x_j)$.
- Recall that we were able to derive the conclusion of the representer theorem using strong duality for SVMs.

Remarks

- It's much easier to compute the kernel $k(x, y)$ instead of the inner product.
- The kernel $k(x, y)$, to some extent, represents a similarity score between two data points.
- The predictor function is basically assigning a value to the new value base on the values near it.
- We are almost guaranteed to overfit on training data (we have N data points and N parameters), regularization is very important.

Some Math that was skipped

- Pre-Hilbert Space
- Hilbert Space
- Orthogonality

Inner Product Space (or "Pre-Hilbert" Spaces)

An **inner product space** (over reals) is a vector space $\mathcal{V}$ with an **inner product**, which is a mapping

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$$

that has the following properties: $\forall x, y, z \in \mathcal{V}$ and $a, b \in \mathbb{R}$:

- Symmetry: $\langle x, y \rangle = \langle y, x \rangle$
- Linearity: $\langle ax + by, z \rangle = a \langle x, z \rangle + b \langle y, z \rangle$
- Positive-definiteness: $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \iff x = 0$.

To show a function $\langle \cdot, \cdot \rangle$ is an inner product, we need to check the above conditions.

Hilbert Space

- A pre-Hilbert space is a vector space equipped with an inner product.
- We need an additional technical condition for Hilbert space: **completeness**.
- A space is **complete** if all Cauchy sequences in the space converge to a point in the space.

Definition

A **Hilbert space** is a complete inner product space.

Example

Any finite dimensional inner produce space is a Hilbert space.

Orthogonality (Definitions)

Definition

Two vectors are **orthogonal** if $\langle x, x' \rangle = 0$. We denote this by $x \perp x'$.

Definition

x is orthogonal to a set S, i.e. $x \perp S$, if $x \perp s$ for all $x \in S$.

Pythagorean Theorem

Theorem (Pythagorean Theorem)

If $x \perp x'$, then $\|x + x'\|^2 = \|x\|^2 + \|x'\|^2$.

Proof.

We have

$$\begin{aligned} \|x + x'\|^2 &= \langle x + x', x + x' \rangle \\ &= \langle x, x \rangle + \langle x, x' \rangle + \langle x', x \rangle + \langle x', x' \rangle \\ &= \|x\|^2 + \|x'\|^2. \end{aligned}$$

$\square$

References

- DS-GA 1003 Machine Learning Spring 2021