4. Kernel

4.1 What's kernel trick

For intuition, **kernel trick** is just a map function.

> k(x, x'): $R^d \times R^d \to R$

The reason we need that is for convenience. At some extent, it's the same reason we need multiplication.

All multiplication can be replaced by addition, but we still want it. Like **1+2+3+4+5=5*3(Gauss Summation)**, with multiplication, we don't need explicitly sum each elements. All we need is the final result.

And that's also true about **Kernel trick**.

Suppose we have a kernel of degree 2: $k(x, x') = (x^T x' + 1),\ \text{where } x, x' \in R^2$

And we have a mapping function $\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2)$.

You will find that $(x^T x' + 1)^2 = \phi(x)^T \phi(x')$.

Obviously, it means that

> 1. we don't need to **explicitly** get the **mapping** function.
> 2. we don't need to waste that much resources on **computation**.

Now, back to machine learning.

4.2 Kernel Methods

In previous, we used the feature of data to make prediction. But we can still make prediction with the relationship of data points.

Hypothesis

$$h_\theta(x) = \theta_0 + \theta_1 similarity(x, x^1) + \theta_2 similarity(x, x^2) + \ldots + \theta_m similarity(x, x^m)$$

Compared with the hypothesis we have before:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_d x_d$$

Actually, it's like we change each data point x into a new x, , consisted of the relationship with others, with function similarity. And the following thing is just the same, the loss function, the gradient descent and so on, since we only changed the data point.

As for function similarity, one typical function is $similarity(x, x') = exp(-\frac{||x-x'||^2}{2\sigma^2})$. That's a kernel. And of course it can be replaced by other kernel functions.

One thing needs to notice is that we don't actually need the whole data point, we can just pick a few landmarks.

Mercer's theorem

Only when the $m \times m$ **Gram matrix** K whose elements are $k(x^i, x^j), 1 \leq i, j \leq m$ are **positive definite** for all choices of the set $x^i : 1 \leq i \leq m$, this kernel function is a **valid** function.

4.3 Kernelize Linear regression

Given

> Dataset D: $((x^i, y^i)|1 \leq i \leq m,\ x^i \in R^d,\ y^i \in R)$
>
> Basis function ϕ: $R^d \rightarrow R^D$
>
> Hypothesis: $h_\theta(x)$: $\theta^T \phi(x)$

The Loss function we already knew:

I'll use ϕ to represent $\phi(x)$, more convenient to write and contrast

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\theta^T \phi^i - y^i)^2 + \frac{\lambda}{2m} \theta^T \theta$$

And we can have the derivative and thus get θ.

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{1}{m} \sum_{i=1}^{m} (\theta^T \phi^i - y^i) \phi^i + \frac{\lambda}{m} \theta = 0$$

$$\theta = -\frac{1}{\lambda} \sum_{i=1}^{m} (\theta^T \phi^i - y^i) \phi^i$$

Let $a^i = -\frac{1}{\lambda}(\theta^T \phi^i - y^i)$, we have:

$$\theta = \phi^T a$$

The Loss function when substituting θ with $\phi^T a$:

$$J(\theta) = \frac{1}{2m} (\phi\theta - y)^T (\phi\theta - y) + \frac{\lambda}{2m} \theta^T \theta$$

$$J(a) = \frac{1}{2m} (\phi\phi^T a - y)^T (\phi\phi^T a - y) + \frac{\lambda}{2m} (\phi^T a)^T (\phi^T a)$$

Let $K = \phi\phi^T$, we have

$$J(a) = \frac{1}{2m} (Ka - y)^T (Ka - y) + \frac{\lambda}{2m} a^T K a$$

$$a = (K + \lambda I)^{-1} y$$

This loss function says that our new dataset now is K. And our parameter is a.

That implies the procedure of kernelized linear regression. I'll record it in **4.1.4** .

4.4 Classic regression VS Kernel regression

Classic regression:

> 1. With basis function ϕ: $R^d \rightarrow R^D$, making feature augmentation and change the each data into R^D.
> 2. The closed form: $\theta = (\phi^T\phi + \lambda I)^{-1}\phi^T y$
> 3. Prediction: $\theta^T\phi(x)$
> 4. θ **corresponds to each feature in new space.**

Kernel regression:

> 1. With kernel function k: $R^d \times R^d \rightarrow R$, implicitly make feature augmentation and change the original dataset into Gram Matrix with size $m \times m$.
> 2. The closed form: $a = (K + \lambda I)^{-1}y$
> 3. Prediction: $a^T k(x)$
> 4. a **corresponds to each data example in new datasets.**