Lecture 10: Kernel methods

wbg231

December 2022

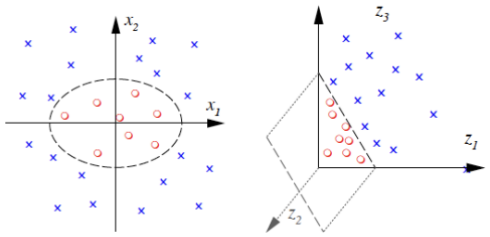
1 feature maps

- explicit feature map for linear model let $\phi : X \to \mathbb{R}^d$ be a feature and our hypothsis space be the affine functions on that feature space

$$\mathcal{F} - \{x \to w^t \phi(x) + b | w \in \mathbb{R}^D, b \in \mathbb{R}\}$$

- the idea is that we identify some feature space, which we can map our features onto to make the classes linearly seperable

$$\Phi : R^2 \to R^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{(2)} x_1 x_2, x_2^2)$$



- adding more feature, allows our linear models to become more expressive
- we can either explicitly specicfy feature but this requires domain knowledge or think of our features as building blocks that can be put together
- we can not do this naively however as adding more features to our representation leads our data matrix to become very large causes overfitting and memory issues
- we can try to avoid overfitting with regularization and cross validation
- we can try to solve computational or memory issues with kernel methods

kernel methods

- svm objective with an explicit feature map can be expressed as

$$j(w) = \frac{1}{2}||w||_2^2 + \frac{c}{n}\sum_{i=1}^{n} max(0, 1 - y_i w^t \phi(x_i))$$

- this yields the following dual problem

$$max \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j < \phi(x_j), \phi(x_i) >$$

$$st \quad \sum_i \alpha_i y_i = 0, \alpha_i \in [0, \frac{c}{n}]$$

- note that $\phi(x)$ our feature map only shows up in inner products in both our training and and observations
- note that we can calculate the inner product of our data tansformed in the feature space with out actually touching the feature space this means our computational cost goes down dramatically

kernel function

- we define a kernel function as the inner product of two feature maps over some hilbert space

$$k(x, x') =< \phi(x), \phi(x') >$$

- a method is kernelized if every feature vector $\phi(x)$ only appears inside an inner product with another feature vector $\phi(x')$. this must hold for both the optimization problem and prediction function
- call the kernel matrix a matrix such that each enterry is the kernel of the coresponding enteries in the orignal data matrix that is $K_{i,j} = k(x_i, x_j)$
- the kernel matrix summarizes all information across the training inputs $x_1 \cdots x_n$ that are required to solve teh optimization problem
- when we kernelized an algorithm we can swap out the inner product, for a new kernel function that may correspond to a very high feature space
- there is an upfront cost of computating the kernel matrix of $O(d)$ but once we have computer that kernel the computaitonal cost of prediction depends on the number of data points not the sie of the feature space
- think of kernel as similarity scores in some hilbert space
- a symmetric function is positive definite if the kernel matrix generated by that fucntion is positive semi defininte for all potnetial sets of inputs

- note that a symmetric function can be expressed as an inner product of some feature map $\iff$ if the kernel function is PD
- can modify old kernel to make new ones
- linear kernel is just $K(x, x') = x^t x'$
- quadratic kernel $< \phi(x), \phi(x') >=< x, x' > + < x, x' >^2$

representer theorem

- notice that in svm our optimal parameter is

$$\phi^* = \sum_i \alpha_i^* y_i x_i = \sum_i \gamma_i x_i \iff w^* \in span(X)$$

- ridge regression has the closed form solution $(X^T X + \lambda I)^{-1} X^t y$
- note that we can write $sw^* = X^T(\frac{1}{\lambda} y - \frac{1}{\lambda} X w^*) = X^t \alpha^* \iff w^* \in span(X)$
- so note that our orignal ridge regression

$$w^* = argmin_{w \in \mathbb{R}^d} \sum_{i=1}^{n} (w^t x_i + y_i)^2 + \lambda ||w||^2$$

given the fact that w" $\in span(X)$, we cna just mnimize over that span(x) that is

$$w^* = argmin_{w \in span(X)} \sum_{i=1}^{n} (w^t x_i + y_i)^2 + \lambda ||w||^2 = argmin_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} ((X^T \alpha)^t x_i - y_i)^2 + \lambda ||X^T \alpha||_2^2$$

- so we have taken our search space from $\mathbb{R}^d$ to $\mathbb{R}^n$

General objective function over linear hypothesis space

- given w, x_1.
$cdotsx_n \in \mathcal{H}$ from some hilbert space
- where the norm $||*||$ is the norm of the corresponding inner product space
- R is a regularization temr
- and L is a loss term

- we can write our generalized linear objective as

$$j(w) = min_{w \in \mathbb{H}} R(||w||) + L(< w, x_1 > \cdots < w, x_n >)$$

- Representer theorem any function that can fit into the form

$$j(w) = min_{w \in \mathbb{H}} R(||w||) + L(< w, x_1 > \cdots < w, x_n >)$$

  ie generalized linear objective will have a minimizer w such that $w \in span(X)$

- if we define$K = X^T X$ and write $w = \sum_i^n \alpha_i x_i$ than we can re-write our objective function in terms of our kernel matrix K as

$$J_0(\alpha) = R(|| \sum_{i=1}^{n} \alpha_i x_i ||) + L(s(\sum_i \alpha_i x_i)) = R(\sqrt{\alpha^t K \alpha}) + L(K\alpha)$$

- which we can mnimize over $\alpha \in \mathbb{R}^n$ and our prediction function can be expressed as $\hat{f} =< w^*, x >=< \sum_{i=1}^n \alpha_i x_i, x_j >= \sum_{i=1}^n < x_i, x_j >= \alpha K$

- by the recomender theorem all norm-regualrized mdoels can be kernalized

- so this tells us that all info we need about our data X is Summarized in it's gram matrix