Lecture 1 intro to statistical learning theory

wbg231

December 2022

1 introduction

- for stuff before the midterm i am going to review it pretty quickly unless it is really important

decision theory definitions

- a prediction function gets input $x \in X$ as inputs and produces $a \in A$

$$f : X \to a$$

- a loss function evaluates an action in the context of the outcome y that is

$$\ell : A \times y \to \mathbb{R}$$

- risk over a prediction function $f : X \to A$ is

$$R(f) = E_{(x,y)\sim P_{x,y}}[\ell(f(x), y)]$$

- we can not compute this in practice since we do not know the true data generating process
- the bayes prediction function is the minimal risk prediction function that is
$$f^* \in argmin_i R(f)$$
- since we can not compute risk we can use the empirical risk of a function $f : X \to A$ with respect to dataset D is

$$\hat{R}_n(f) = \frac{1}{n}\Sigma_{i=1}^n \ell(f(x_i), y_i)$$

- in other words it is the mean loss over our training data if we predict using that function

- the empirical risk minimizer is the the function

$$\hat{f} \in argmin_f \hat{R}_n(f)$$

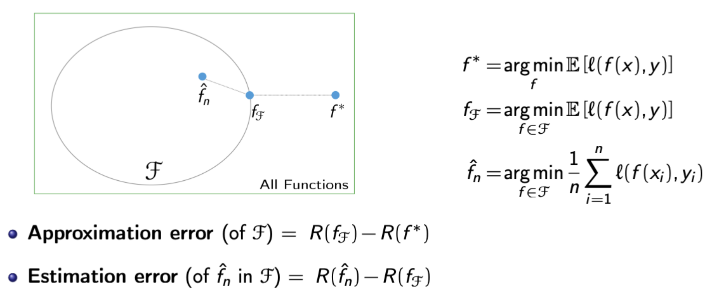
- that is the function that could get the minimal risk on our training set
- in many cases unconstrained ERM will just memorize the training set,
- so to improve generalization we can use constrained ERM, that is instead of minimzed risk of all prediction functions we constrain our search space to a set of functions called the hypothesis space
- so we can get constrained find our constrained empirical risk minizer as

$$\hat{f}_n \in argmin_{f \in \mathcal{F}} \hat{R}_n(f)$$

  where $\mathcal{F}$ is our hypothesis space

- a risk minimizer in $\mathcal{F}$ is

$$f^*_{\mathcal{F}} \in argmin_{f \in \mathcal{F}} E[\ell(x), y]$$

- 

- as we can see our Approximation error is what we lose by specifying a hypothesis class
- our estimation error is the diference in how well our data could be fit by any function veruss the best in our hypothesis class
- excess risk is defined as

$$R(f) - R(f*)$$

  that is the diference in risk between our learned function and the bayes optimal one

- we can wrote excess risk as the sum of Approximation error and estimation error, so there is a trade off between the two

- a larger hypothesis space means smaller Approximation error (Approximation error is a non random variable it is a function of our hypothesis space)
- estimation error goes up as our hypothesis space gets more complex. it is a random variable due as a function of our data
- in practice we can not simply find and argmin in most cases so we call the optimization error the difference between the true empirical risk minimizer and what our optimization learned in practice
- so overall we can think of excess risk = optimization error + estimation error + Approximation error
- we can not observe this in practice