1 A General Framework of Statistical Learning

Recall $f : X \rightarrow Y$ is a prediction function from population $X$ to label set $Y$. Suppose there is a probability distribution over $X \times Y$, and its probability density function (pdf) is denoted by $p$. We define loss function $\ell(f(x), y)$, which measures the unfitness of $f$ on instance $(x, y)$. An example is mean-squared-loss $\ell(f(x), y) = (f(x) - y)^2$; another is 0-1 loss $\ell(f(x), y) = 1_{f(x) \neq y}$.

A general framework to learn $f$ is to minimize its expected loss i.e. $E[\ell(f(x), y)] = \int_{(x,y) \in X \times Y} \ell(f(x), y) \cdot p(x, y) \, dx \, dy \quad (1)$

Note $E[\ell(f(x), y)]$ does not depend on any particular instance as the expectation is taken over the randomness of $(x, y)$. So in the rest discussions, we will denote it as $E[\ell(f)]$.

Ideally, we want to learn a function $f$ that minimizes $E[\ell(f)]$, but this is often impossible as it requires one to have all instances in $X \times Y$ (in order to calculate expected loss).

So in practice, we often learn a function that minimizes empirical loss, whose calculation only requires a sample $S \subseteq X \times Y$. The empirical loss on sample $S$ is defined as $L(f; S) = \frac{1}{|S|} \sum_{(x,y) \in S} \ell(f(x), y). \quad (2)$

[Question] Is $L(f; S)$ a good approximation of $E[\ell(f)]$? Yes. Here is a statistical justification:

$E[L(f; S)] = E[\ell(f)], \quad (3)$

where the left expectation is taken over randomness of $(x, y) \in S$.

[Exercise] Prove (3).

Ideally, if we have sufficient instances in $S$, we can learn a good $f$ by just minimizing empirical loss. But in reality instances may be insufficient, then we may learn a biased $f$ (over-fitting).

So in practice, we often add a regularization term to empirical loss i.e.

$L(f; S) = \frac{1}{|S|} \sum_{(x,y) \in S} \ell(f(x), y) + \lambda \cdot \Omega(f), \quad (4)$

where $\Omega()$ is a function measuring complexity of $f$, and $\lambda$ is a regularization coefficient balancing the importance of function fitness and complexity. Larger $\lambda$ means learning a simpler model, which may fit instances less accurately but suffers less overfitting.

When training sample $S$ is given, its size $|S|$ is a constant. Thus we may omit it when learning $f$ (by minimizing (2) or (4)). So one often minimizes $L(f; S) = \sum_{(x,y) \in S} \ell(f(x), y) + \lambda \cdot \Omega(f). \quad (5)$

[Remark] There are different loss functions and regularization terms; they define different machine learning algorithms. Most learning algorithms, but not all, are describable by (5).