

LECTURE 2: RISK MINIMIZATION

In this lecture we introduce the concepts of empirical risk minimization, overfitting, model complexity and regularization.

1 Empirical Risk Minimization

Given a loss function $\ell(\cdot, \cdot)$, the risk $R(h)$ is not computable as $P_{X,Y}$ is unknown. Thus we may not be able to directly minimize $R(h)$ to obtain some predictor. Fortunately we are provided with the training data $\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ which represents the underlying distribution $P_{X,Y}$.

Instead of minimizing $R(h) = \mathbb{E}_{X,Y}[\ell(Y, h(X))]$, one may replace $P_{X,Y}$ by its empirical distribution and thus obtain the following minimization problem:

$$\hat{h}_n = \arg \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, h(x_i))$$

which we call *empirical risk minimization* (ERM). Furthermore, we also define the *empirical risk* $\hat{R}_n(h)$ as

$$\hat{R}_n(h) := \frac{1}{n} \sum_{i=1}^n \ell(y_i, h(x_i)).$$

Because under some conditions $\hat{R}_n(h) \rightarrow_p R(h)$ by the law of large numbers, the usage of ERM is at least partially justified.

ERM covers many popular methods and is widely used in practice. For example, if we take $\mathcal{H} = \{h(x) : h(x) = \theta^T x, \forall \theta \in \mathbb{R}^p\}$ and $\ell(y, p) = (y - p)^2$, then ERM becomes the well-known least squares estimation. The celebrated *maximum likelihood estimation* (MLE) is also a special case of ERM where the loss function is taken to be the negative log-likelihood function. Example: in binary classification $(x_1, y_1), \dots, (x_n, y_n)$ where $y_i \in \{-1, 1\}$ and $\mathcal{H} = \{h(x) : h(x) = \theta^T x, \forall \theta \in \mathbb{R}^p\}$, the logistic regression is computed by minimizing the logistic loss:

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i \theta^T x_i))$$

which is equivalent to MLE.

2 Overfitting

ERM works by minimizing the empirical risk $\hat{R}_n(h)$, while the goal of learning is to obtain a predictor with a small risk $R(h)$. Although under certain conditions the former will converge to the latter as $n \rightarrow \infty$, in practice we always have a finite sample and as a result, there might be a large discrepancy between those two targets, especially when \mathcal{H} is large and n is small. *Overfitting* refers to the situation where we have a small empirical risk but still a relatively large true risk.

Consider the following example. Let $\ell(y, p) = (y - p)^2$ and we obtain the predictor \hat{h} by ERM:

$$\hat{h} = \arg \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n (y_i - h(x_i))^2.$$

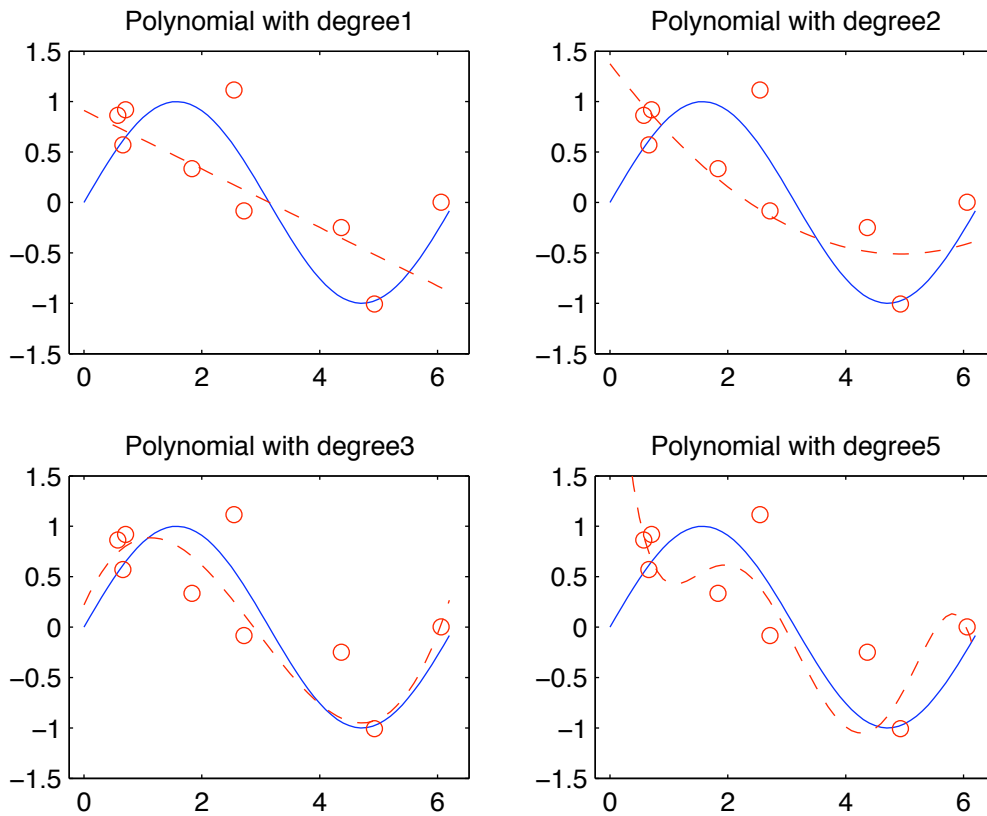


Figure 1: Overfitting of polynomial regression. The true signal function (blue line) is $h^*(x) = \sin(x)$, and the function is fitted using 10 training examples (red dots). \mathcal{P}_1 and \mathcal{P}_2 show a lack of fitting (underfitting) and \mathcal{P}_5 is overfitting.

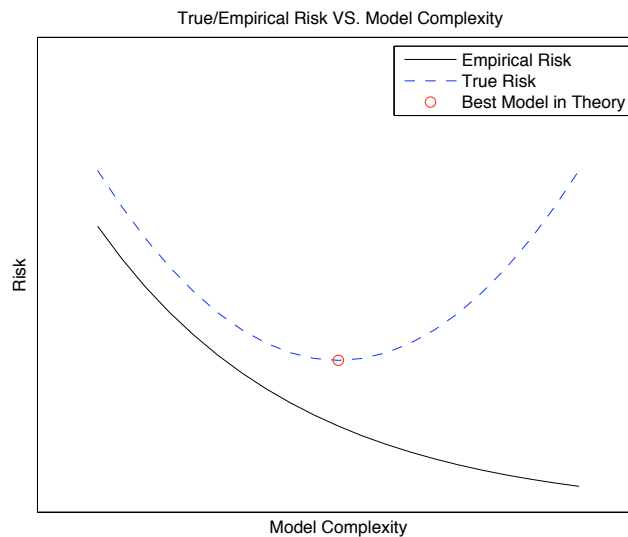


Figure 2: True/empirical risk vs. model complexity

Figure 1 shows the case where \mathcal{H} is taken to be $\mathcal{P}_1, \mathcal{P}_2, \dots$, where \mathcal{P}_k is the set of all polynomial functions with order up to k .

We can see that when $\mathcal{H} = \mathcal{P}_3$ the fitted predictor will have a small risk (close to the true signal $\sin(x)$). Taking \mathcal{P}_k with larger k values as the hypothesis space can clearly improve its fitting with respect to the 10 observations (red dots), but this does not necessarily reduce the true risk as it overfits the training data. Learning is more about generalization than memorization.

3 Controlling Model Complexity

Overfitting is mainly caused by the fact that the hypothesis space \mathcal{H} is too large for the sample size n . Clearly the complexity of the hypothesis space \mathcal{H} (i.e the size of \mathcal{H}) we can afford depends on the amount of training data we have. For a given training dataset, the relationship between the true risk, the empirical risk and model complexity can be best illustrated as in Figure 2.

One way to avoid overfitting is to choose \mathcal{H} so that it is appropriate for the sample size. There are many ways to control the model complexity, and they are in fact quite similar in spirit. Here we list two commonly used approaches:

1. Take $\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_n, \dots$ to be a sequence of increasing sized spaces. For example, one typically has $\mathcal{H}_k \subset \mathcal{H}_{k+1}$ and $\bigcup \mathcal{H}_k = \mathcal{H}$. Given the training data \mathcal{D}_n one finds \hat{h}_n by minimizing

$$\hat{h}_n = \arg \min_{h \in \mathcal{H}_n} \hat{R}_n(h).$$

This covers the method of *Sieves* and *structural risk minimization* (SRM).

2. Define a penalty function $\Omega : \mathcal{H} \mapsto \mathbb{R}^+$ and find \hat{h}_n by the following optimization procedure:

$$\hat{h}_n = \arg \min_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \ell(y_i, h(x_i)) + \lambda_n \Omega(h)$$

where $\lambda_n > 0$ balances the trade-off between goodness-of-fit and model complexity. This is also known as the *penalized empirical risk minimization*.

In practice we often need to select \mathcal{H}_n or λ_n based on the training data to achieve a good balance between goodness-of-fit and model complexity.

Consider the following regression problem: let $\mathcal{H} = \{h(x) : h(x) = \theta^T x, \forall \theta \in \mathbb{R}^p\}$ and we are trying to find an estimator $\hat{\theta}$ which minimizes the risk $\mathbb{E}_{X,Y} (Y - \theta^T X)^2$. For the first approach, we could define a sequence of increasing constants $0 \leq \eta_1 \leq \eta_2 \leq \dots \leq \eta_k \leq \dots$ and define $\mathcal{H}_k = \{h(x) : h(x) = \theta^T x, \theta^T \theta \leq \eta_k\}$. For the second approach we define $\Omega(h) = \theta^T \theta$. Then it is well-known from optimization that those two approaches become mathematically equivalent (i.e. for any η_k there exists a λ such that those two optimization problems have the same solution).