

Lecture 12 — February 20

Scribe: Yasser Ganjisaffar

Lecturer: Deva Ramanan

Note: These lecture notes are still rough, and have only have been mildly proofread.

12.1 Learning Theory

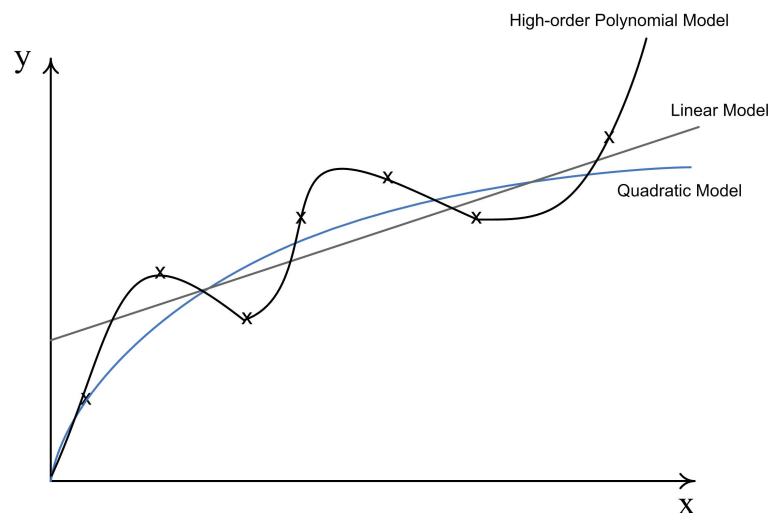
Consider task of classification,

- When will a particular algorithm work?
- How much data is needed?
- What is the correct model class?

Learning theory is an attempt to answer these questions.

12.1.1 Intuition: Bias versus Variance

Consider the following experiment - we'll try to fit a linear regression model to a dataset, similar to the one from HW3, that is generated from a quadratic model $y = ax + bx^2$.



We can imagine fitting a linear regression model, a quadratic regression model, and a high-order (say 10^{th} order) polynomial model. We could expect the linear and quadratic model to have poor performance on test data, but for different reasons

Bias : Linear model has a bias that won't disappear even with infinite training data.

Variance : High-order polynomial model has larger variance. We're more likely to overfit to a specific training data. You probably saw this in your cross-validation experiments for selecting polynomial order.

One of the fundamental problems in machine learning is model selection. A model that is too complex could overfit the training data and a model that is too simple could be a bad approximation of the function that we are trying to estimate. We can formalize these notions of bias and variance, as well as getting a firmer handle on our initial questions, below.

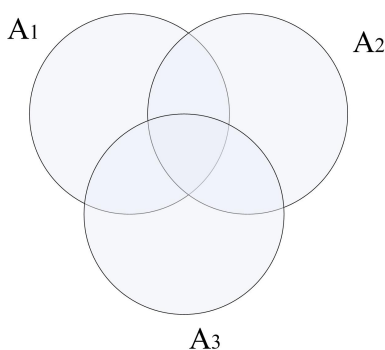
12.1.2 Background

Union Bound

In probability theory, *union bound* says that for any finite or countable set of events, the probability that at least one of the events happens is no greater than the sum of the probabilities of the individual events. An event A can be thought of as a set of outcomes from some sample space Ω . For example, A could be the event that a dice roll yields a 1, 2, or 6. Formally, for a countable set of events A_1, \dots, A_k , we have:

$$P(A_1 \cup A_2 \cup \dots \cup A_k) \leq P(A_1) + P(A_2) + \dots + P(A_k) \quad (12.1)$$

Intuitively, the above inequality can be shown to be true using Venn diagrams: We can



visualize an event as a region in a sample space - above, we use circles. The probability of the event is proportional to the size of the set (or circle). The probability of the union of the 3 events is the total area of the circles. We can calculate that by adding up the area of each circle and subtracting the overlapped regions that are double counted.

Chernoff Bound (Hoeffding inequality)

Let z_1, \dots, z_m be independent random variables and identically distributed from $\text{Bernoulli}(\phi)$ and let $\hat{\phi} = \frac{1}{m} \sum_{i=1}^m z_i$. Note that $\hat{\phi}$ is a random variable itself, because it is a function of a random variable. Therefore, we can take expectations of $\hat{\phi}$. In particular, we can show that $\text{var}(\hat{\phi}) = \frac{1}{m} \phi(1 - \phi)$, or its variance decreases linearly with m . The Chernoff Bound intuitively states that the probability that a random variable lies far from its mean is bounded by (some function of) its variance.

$$p(|\hat{\phi} - \phi| > \gamma) \leq 2 \exp(-2\gamma^2 m) \quad (12.2)$$

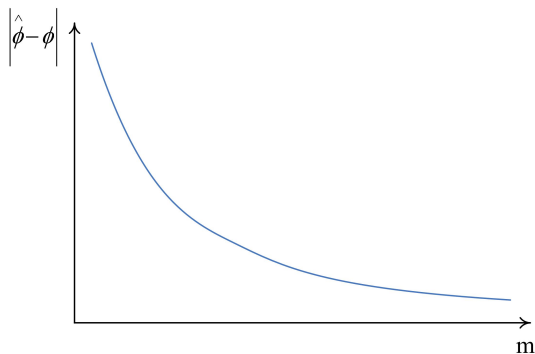


Figure 12.1. In the above figure, the y-axis should read $2 \exp(-2\gamma^2 m)$

12.1.3 Setup

Training set $S = \{(x^{(i)}, y^{(i)}) : i = 1 \dots m\}$. We are assuming these samples are independent and identically distributed (i.i.d.) from the some joint distribution $D = p(x, y)$.

We've already considered such joint distributions. For class-conditional Gaussians, we modelled $p(x, y) = p(y)p(x|y)$ as

$$\begin{aligned} p(y) &= \text{Bernoulli}(\phi) \\ p(x|y) &= \text{Gaussian}(\mu_y, \Sigma_y) \end{aligned}$$

In the following, we will prove bounds that hold for *any* joint distribution D . Let h be a particular classifier from an family of possible classifiers $h \in \mathcal{H}$. We can compute the number of mistakes h makes on the set S .

Empirical Error / Empirical risk / Empirical loss:

$$\hat{\varepsilon}(h) = \frac{1}{m} \sum_i I\{h(x^{(i)}) \neq y^{(i)}\} \quad (12.3)$$

Note that: $I\{x\} = \begin{cases} 1, & \text{if } x = \text{true} \\ 0, & \text{otherwise} \end{cases}$

Generalization error:

$$\varepsilon(h) = \mathbb{P}_{x,y \sim D} (h(x) \neq y(x)) \quad (12.4)$$

As we see more samples, $\hat{\varepsilon}(h)$ converges to $\varepsilon(h)$:

$$\hat{\varepsilon}(h) \xrightarrow{m \rightarrow \infty} \varepsilon(h) \quad (12.5)$$

Note that ideally, we would like to select a classifier h that minimizes generalization error. In practice, that error is not computable, so we often instead select a classifier that minimizes empirical error.

Hypothesis class \mathcal{H}

- infinite \mathcal{H} (e.g., logistic regression).
- finite \mathcal{H} (e.g., decision trees for boolean inputs).

Empirical risk minimization:

We select the hypothesis that minimizes our empirical loss:

$$\hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} \hat{\varepsilon}(h) \quad (12.6)$$

We haven't been doing this. But, informally speaking, algorithms like logistic regression and support vector machines try to approximately do this. We'll talk more about why optimizing empirical loss is hard later.

12.1.4 Finite \mathcal{H}

Consider $h_i \in \mathcal{H}$ and some $D = p(x, y)$, let $z_j = I\{h_i(x^{(j)}) \neq y^{(j)}\}$. z_j are independent and identically distributed from $\text{Bernoulli}(\varepsilon(h_i))$:

$$\hat{\varepsilon}(h_i) = \frac{1}{m} \sum_{j=1}^m z_j \quad (12.7)$$

Using Chernoff Bound:

$$p(|\hat{\varepsilon}(h_i) - \varepsilon(h_i)| > \gamma) \leq 2 \exp(-2\gamma^2 m) \quad (12.8)$$

Define $A_i = \{ \text{Bad hypothesis} \} = \{ |\hat{\varepsilon}(h_i) - \varepsilon(h_i)| > \gamma \}$. Then we would have:

$$p(A_i) \leq 2 \exp(-2\gamma^2 m) \quad (12.9)$$

$$\underbrace{p(\exists h \in \mathcal{H} : |\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma)}_{\delta} = p(A_1 \cup A_2 \cup \dots \cup A_k) \text{ where } k = |\mathcal{H}| \quad (12.10)$$

Applying the union bound:

$$\delta \leq \sum_{j=1}^k p(A_j) \quad (12.11)$$

$$\delta \leq 2k \exp(-2\gamma^2 m) \quad (12.12)$$

Inequality 12.12 is called *uniform convergence result*.

Main quantities:

- m = number of examples.
- γ = discrepancy between empirical and generalization error.
- δ = probability of error (due to unlucky dataset S).

Solving inequality 12.12 for m would result (for a fixed γ and δ):

$$m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta} \quad (12.13)$$

If inequality 12.13 is true then with probability $1 - \delta$,

$$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma, \forall h \in \mathcal{H} \quad (12.14)$$

There is always a chance that it is impossible to arbitrarily bound the error of h because of a highly abnormal training set. Thus, we want the algorithm generating h to be *probably approximately correct* (PAC). “Probably” comes from the δ condition (that we may have encountered an unlucky set of training samples S), and “approximately” comes from γ discrepancy between the generalization and empirical error.

Equation 12.13 allows us to ask questions about *sample complexity* - how many training examples are needed to ensure a particular performance guarantee on unseen data (with high

probability)?

Example:

Lets assume: $k = 1000$ and $\hat{\varepsilon}(h) = 0.05$. We choose $\delta = 0.01$ and $\gamma = 0.1$. If $m > 600$, then with 99% probability, generalization error is less than 0.15. We can also look at how the number of samples varies with the size of the hypothesis space k . If we increase k by a factor of 10, we need only double m to obtain a similar guarantee.

Solving inequality 12.12 for γ would result:

$$\gamma \leq \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}} \quad (12.15)$$

Combining this with inequality 12.14 would result:

$$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}} \quad (12.16)$$

Inequality 12.16 is true with probability $1 - \delta$ and is often called the PAC bound.

Now imagine we could select the best possible classifier:

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \varepsilon(h) \quad (12.17)$$

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \hat{\varepsilon}(h) \quad (12.18)$$

$$\varepsilon(\hat{h}) \leq \hat{\varepsilon}(\hat{h}) + \gamma \quad (12.19)$$

$$\leq \hat{\varepsilon}(h^*) + \gamma \quad (12.20)$$

$$\leq \varepsilon(h^*) + 2\gamma \quad (12.21)$$

$$(12.22)$$

Theorem 12.1.

$$\varepsilon(\hat{h}) \leq \underbrace{\min_{h \in \mathcal{H}} \varepsilon(h)}_{\varepsilon(h^*)} + 2\sqrt{\frac{1}{2m} \log \frac{2k}{\delta}} \quad \text{holds with probability at least } 1 - \delta \quad (12.23)$$

Assume we want to minimize the generalization error for a fixed δ and m , but varying the size of the hypothesis class $k = |\mathcal{H}|$. We want to minimize the two terms on the RHS. The first term can be minimized by selecting a large k , while the second term will be minimized by selecting a small k . We can think of the first term as encoding the error due to our model bias (which can be minimized by considering a large class of models), and the second term encodes the error to the model variance (which can be minimized by considering a small class of models).

Often, this is written as:

$$\varepsilon(h) \leq \hat{\varepsilon}(h) + \sqrt{\frac{1}{2m} \log \frac{k}{\delta}} \quad \text{holds for all } h \text{ with probability at least } 1 - \delta \quad (12.24)$$

(12.25)