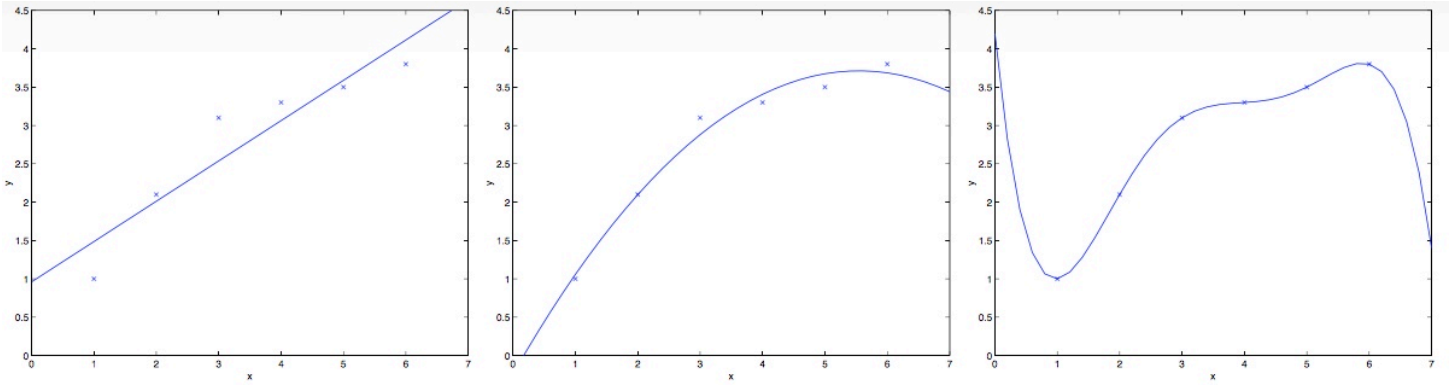
# CS229 Note Lecture 09

## Bias / Variance



underfitting - high bias

overfitting - high variance

## Linear Classification

Assume linear problem as

$$
\begin{aligned}
h_\theta(x) &= g(\theta^T x) \\
g &= 1\{z \geq 0\} && (y \in \{0, 1\}) \\
S &= \{(x^{(i)}, y^{(i)})\}_{i=1}^{m} && (x^{(i)}, y^{(i)}) \sim_{\text{IID}} D
\end{aligned}
$$

Define **training error** of $h_\theta$ as

$$
\hat{\epsilon}(h_\theta) = \hat{\epsilon}_S(h_\theta) = \frac{1}{m} \sum_{i=1}^{m} 1\{h_\theta(x^{(i)} \neq y^{(i)}\}
$$

training error also called **empirical risk** or **empirical error**.

# ERM(Empirical Risk Minimization)

Assume

$$
\hat{\theta} = \arg \max_{\theta} \hat{\epsilon}(h_\theta)
$$

For the results we want to prove, it turns out that it will be useful to think of our learning as not

choosing a set of parameters, but as choosing a function.

Define the hypothesis class $\mathrm{H} = \{h_\theta : \theta \in \mathrm{R}^{n+1}\}$ as the class of all hypothesis of linear classifiers, and $h_\theta \in \mathrm{H} : \mathrm{X} \mapsto \{0, 1\}$.

Redefine ERM as

$$\hat{h} = \arg\max_{h \in \mathrm{H}} \hat{\epsilon}_S(h)$$

The ultimate goal is how well it makes generalization, in other work, how well it makes predictions on examples that we haven't seen before. What we really care about is generalization error

$$\epsilon(h) = P_{(x,y) \sim \mathrm{D}}(h(x) \neq y)$$

# Union Bound / Hoeffding's Inequality

## Union Bound

Let $A_1, A_2, \ldots, A_k$ be $k$ event, and these are not necessarily independent, so there's some current distribution over the events $A_1$ through $A_k$. Then

$$P(A_1 \cup A_2 \cup \cdots \cup A_k) \leq P(A_1) + P(A_2) + \cdots + P(A_k)$$

This usually written as an axiom. And in learning theory it's commonly called the **union balance**.

## Hoeffding's Inequality

Let $z_1, z_2, \ldots, z_m$ be $m$ IID Bernoulli($\phi$) random variables, and this means $P(z_i = 1) = \phi$. Define

$$\hat{\phi} = \frac{1}{m} \sum_{i=1}^{m} z_i$$

and let any $\gamma > 0$ be fixed. Then

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

# Uniform Convergence

Let $\mathrm{H} = \{h_1, h_2, \ldots, h_k\}$, what the ERM would do is it takes the training set and it will look at each of these $k$ functions, pick whichever of these functions has the lowest training error

$$\hat{h} = \arg\max_{h \in \mathrm{H}} \hat{\epsilon}_S(h)$$

We are going to prove there is a bound of the generalization error of $\hat{h}$.

Strategy:

1. $\hat{\epsilon} \approx \epsilon$
2. Show bound on $\epsilon(\hat{h})$.

Fix any $h_j \in \mathrm{H}$, define

$$z_i = 1\{h_j(x^{(i)} \neq y^{(i)}\} \in \{0, 1\}$$

So

$$P(z_i = 1) = \epsilon(h_j)$$

therefore the $z_i$'s themselves are IID random variables.

The training error of our hypothesis $h_j$ is

$$\hat{\epsilon}(h_j) = \frac{1}{m}\sum_{i=1}^{m} z_i = \frac{1}{m}\sum_{i=1}^{m} 1\{h_j(x^{(i)} \neq y^{(i)}\}$$

therefore, by the Hoeffding's inequality

$$P(|\epsilon(h_j) - \hat{\epsilon}(h_j)| > \gamma) \leq 2\exp(-2\gamma^2 m)$$

In order to show that the training error will be a good estimate for generalization error not just for this hypothesis $h_j$ but actually for all $k$ hypothesis, let's define a random event $A_j$ that $|\epsilon(h_j) - \hat{\epsilon}(h_j)| > \gamma$, so what we prove previously can be written as

$$P(A_j) \leq 2\exp(-2\gamma^2 m)$$

Thus we have

$$
\begin{aligned}
P(\exists h_j \in \mathrm{H} \mid |\epsilon(h_j) - \hat{\epsilon}(h_j)| > \gamma) &= P(A_1 \cup A_2 \cup \ldots A_k) \\
&\leq \sum_{i=1}^{k} P(A_i) \\
&\leq \sum_{i=1}^{k} 2\exp(-2\gamma^2 m) \\
&= 2k\exp(-2\gamma^2 m)
\end{aligned}
$$

If we subtract both sides from $1$, we find that

$$
\begin{aligned}
P(\neg\exists h_j \in \mathrm{H} \mid |\epsilon(h_j) - \hat{\epsilon}(h_j)| > \gamma) &= P(\forall h_j \in \mathrm{H} \mid |\epsilon(h_j) - \hat{\epsilon}(h_j)| \leq \gamma) \\
&\geq 1 - 2k\exp(-2\gamma^2 m)
\end{aligned}
$$

So what we are shown is that with probability great than $1 - 2k\exp(-2\gamma^2 m)$, $epsilon(h)$ will be within $gamma$ of $\epsilon(h)$ for all $h \in \mathrm{H}$.

And so just to give this result a name, this is called **uniform convergence**.

## Two Other Equivalent Forms

1. Given $\gamma$ and $m$, what is the probability of uniform conversions? Given $\gamma$ and the probability $\delta$, how large a training set size do we need?

   If we set $\delta = 2k\exp(-2\gamma^2 m)$ and solve for $m$, what we find is that there is an equivalent form of this result that so long as our training set assigns

   $$m \geq \frac{1}{2\sigma^2}\log\frac{2k}{\delta}$$

   then with probability at least $1 - \delta$ we have for all $|\epsilon(h_j) - \hat{\epsilon}(h_j)| \leq \gamma$. And just to give this another name called **"sample complexity" bound**.

2. **Error bound**: hold $m$ and $\delta$ fixed and solve for $\gamma$.

   With probability at least $1 - \delta$, we have

   $$\forall h \in \mathrm{H},\ |\epsilon(h_j) - \hat{\epsilon}(h_j)| \leq \sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$$

---

So the second step of the overall proof is the following, let's assume that uniform conversions holds

$$\forall h \in \mathrm{H},\ |\epsilon(h_j) - \hat{\epsilon}(h_j)| \leq \gamma$$

$$\hat{h} = \arg\min_{h \in \mathrm{H}} \hat{\epsilon}(h)$$

Define

$$h^* = \arg\min_{h \in \mathrm{H}} \epsilon(h)$$

so

$$\begin{aligned} \epsilon(\hat{h}) &\leq \hat{\epsilon}(\hat{h}) + \gamma && \text{by (1)} \\ &\leq \hat{\epsilon}(h^*) + \gamma && \text{by (2)} \\ &\leq \hat{\epsilon}(h^*) + 2\gamma && \text{by (3)} \end{aligned}$$

Let's tie all these things together into a theorem.

Theorem: let $|\mathrm{H}| = k$, and let $m$, $\delta$ be fixed, then with probability at least $1 - \delta$ we have

$$\epsilon(\hat{h}) \leq \Big( \underbrace{\min_{h \in \mathrm{H}} \epsilon(h)}_{\epsilon(h^*)} \Big) + 2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$$

To prove this, we set $\gamma = \sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$, we known Equation (1) holds with probability at least $1 - \delta$

which implies Equation (5).

This result helps us to quantify bias variance tradeoff. In particular, if we have some hypothesis class $\mathrm{H}$, considering switching to some new class $\mathrm{H}'$ by having more features($\mathrm{H} \subseteq \mathrm{H}'$), the tradeoff is what if we switch from $\mathrm{H}$ to $\mathrm{H}'$, $\epsilon(h^*)$ will become better. But what we pay for then is that $k$ will increase. This phenomenon called the **bias variance tradeoff**.

Speaking loosely, we can think of $\min_{h \in \mathrm{H}}$ as corresponding to the bias of the learning algorithm, and $\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$ as corresponding to the variance.

# Corollary

Let $|\mathrm{H}| = k$, and let any $\gamma$, $\delta$ be fixed, in order to guarantee

$$\epsilon(\hat{h}) \leq \min_{h \in \mathrm{H}} \epsilon(h) + 2\gamma$$

with probability at least $1 - \delta$, then it suffices that

$$\begin{aligned} m &\geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta} \\ &= O\left(\frac{1}{\gamma^2} \log \frac{k}{\delta}\right) \end{aligned}$$