Structural Risk Minimization

Our goal is to summarize the result of a particular approach termed *structural risk mimimization*, which balances the perfomance of a model against the amount of test data required to guarantee a certain level of performance.

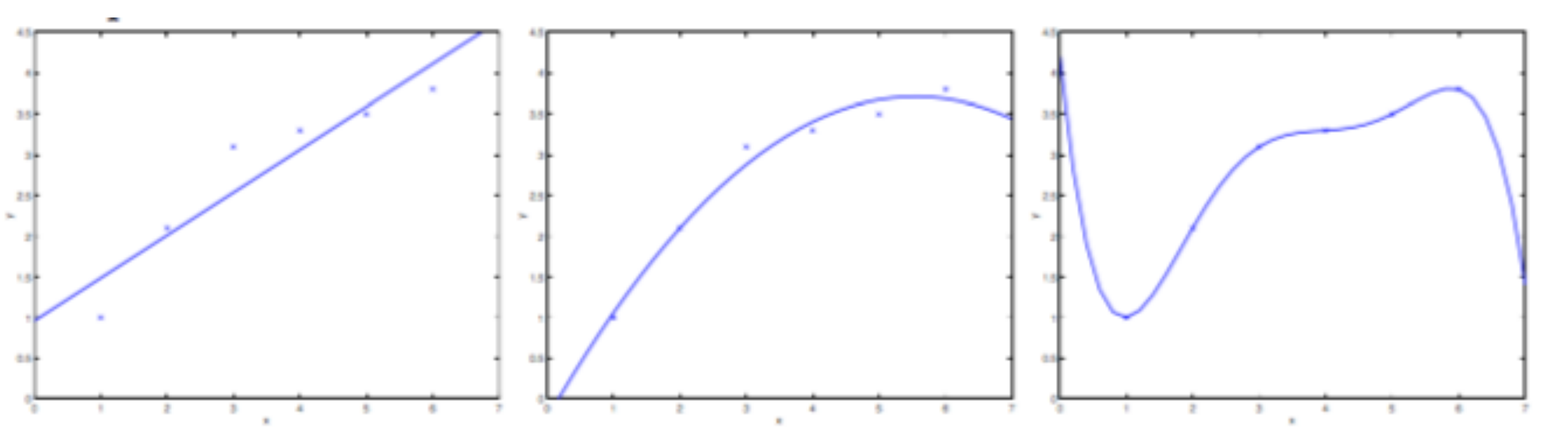
More formally given a hypothesis space$\mathcal{H}$, we want to show that for any instance $h \in \mathcal{H}$ that we can bound the difference between the performance on the test data and the performance on future data, that is where $\hat{\epsilon}(h)$ is the observed error on test data, and $\epsilon(h)$ is the probability of error on the entire data set, that we can bound

$$|\hat{\epsilon}(h) - \epsilon(h)|.$$

We'll do this in three steps. First we show that there is a bound when the size of $\mathcal{H}$ is finite. Next we'll interpret this argument in a finite space that appears very large, for the reason that the result obtained looks a lot like the ultimate result. Finally the introduction of the $VCdimension$ provides an measure that allows the characterization of the power of infinite spaces and thus provide a bound on the performance of infinite $\mathcal{H}$.

The bias variance dilemma

When using a model to fit test data in service of a classification problem there are questions about both the model and the data. If we pick a model that is too weak to capture all the structure of the data, then we maybe missing out in the possible performance we could have by using a more powerful model. On the other hand if the model is too powerful, it may overfit the test data, leading to poor performance on new exemplars.



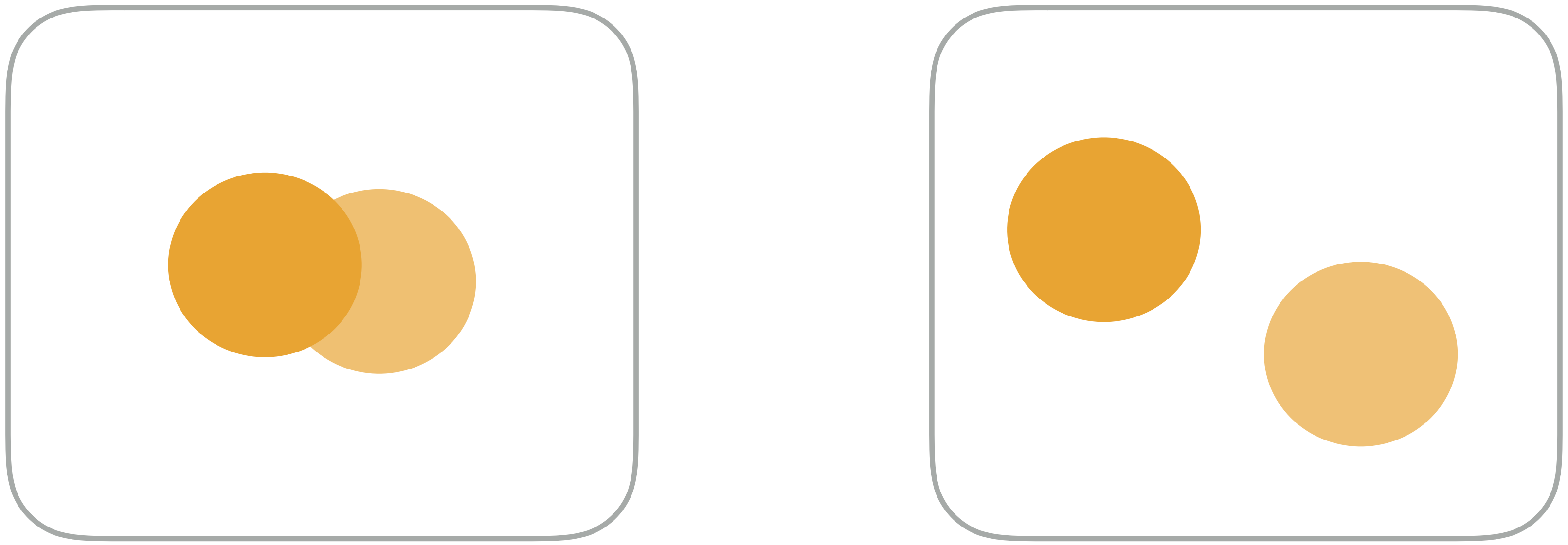
Bias: the generalization error that results from undercutting

Variance: the generalization error that results from overfitting

Two things we'll need

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

I. Union bound



II. Hoeffding inequality

Bernoulli

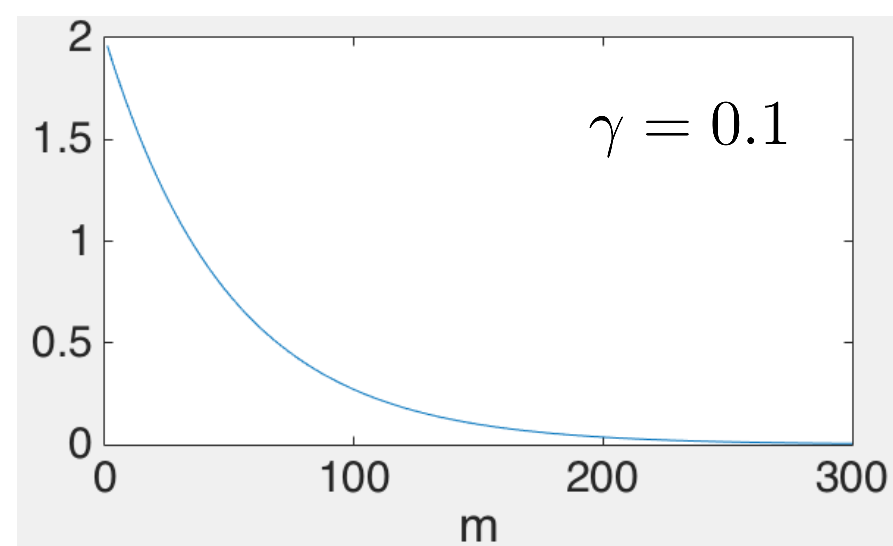
$$\phi = P(Z_i = 1)$$

m samples

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

$$\gamma > 0$$

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



The case of finite $\mathcal{H}$

Consider a training set $S = \{(x^{(i)}, y^{(i)}), i = 1, \ldots, m\}$ where samples are drawn i. i. d. from some probability distribution $\mathcal{D}$.

The *training error* for hypothesis h counts the fraction of errors, i.e.,

$$\hat{\epsilon}(h) = \frac{1}{m} \sum_i 1\{h(x^{(i)}) \neq y^{(i)})\}$$

The *generalization error* denotes the expected number of errors when drawing from $\mathcal{D}$, i.e.,

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

Now we are ready to develop the bound for the case of finite $\mathcal{H}$. Let

$$\mathcal{H} = \{h_1, \ldots, h_k\}.$$

Draw a sample and let the r. v. Z denote whether $h_i(x)$ missclassifies it, i.e.,

$$Z = 1\{h_i(x) \neq y\}$$

Thus empirical traing error is given by

$$\hat{\epsilon}(h_i) = \frac{1}{m} \sum_j Z_j$$

The *Hoeffding inequality* allows the bound between training error and generalization error to be expressed as

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

which can be denoted as an event A_i.

We want this bound to hold for *any* $h_i \in \mathcal{H}$.

$$
\begin{aligned}
P(\exists\, h \in \mathcal{H}.|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(A_1 \cup \cdots \cup A_k) \\
&\leq \sum_{i=1}^{k} P(A_i) && \text{using Union bound} \\
&\leq \sum_{i=1}^{k} 2\exp(-2\gamma^2 m) && \text{using Hoeffding inequality} \\
&= 2k\exp(-2\gamma^2 m)
\end{aligned}
$$

If we subtract both sides from 1, we find that

$$
\begin{aligned}
P(\neg\exists\, h \in \mathcal{H}.|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(\forall h \in \mathcal{H}.|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma) \\
&\geq 1 - 2k\exp(-2\gamma^2 m)
\end{aligned}
$$

Let $\delta = 2k\exp(2\gamma^2 m)$

The result is that “the probability that there is NO empirical hypothesis with an error greater than γ is greater than $1-\delta$

The previous result compared the h for training error
What can we say about the generation error?

Uniform convergence: with probability $1-\delta$

$$|\epsilon(h) - \hat{\epsilon}(h)| \leq \gamma \text{ for all } h \in \mathcal{H}$$

$$\begin{aligned} \varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}(h^*) + \gamma \\ &\leq \varepsilon(h^*) + 2\gamma \end{aligned}$$

Uniform convergence

The hypothesis $\hat{h}$ was chosen to minimize $\hat{\epsilon}(h)$ and in particular $\hat{\epsilon}(\hat{h}) \leq \hat{\epsilon}(h^*)$

Uniform convergence

Now we can ask the crucial question: Given γ and some $\delta > 0$, how large must m be to guarantee that with probability $1 - \delta$ the training error will be within γ of the generalization error?

Set $\delta = 2ke^{-2\gamma^2 m}$ and solve for m:

$$m \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta} \tag{3}$$

If Eq. 3 is satisfied then with probability at least $1-\delta$ the difference between training error and generalization error is $\leq \gamma$ for all the $h \in \mathcal{H}$. Note that this guarantee only logarithmic in k.

Now if you solve Eq. 3 for γ, holding everything else fixed it also follows that:

$$|\hat{\epsilon}(h_i) - \epsilon(h_i)| \leq \sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}$$

Theorem. Let $|\mathcal{H}| = k$, and let any m, δ be fixed. Then with probability at least $1 - \delta$, we have that

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h)\right) + 2\sqrt{\frac{1}{2m} \log \frac{2k}{\delta}}.$$

This is proved by letting γ equal the $\sqrt{\cdot}$ term, using our previous argument that uniform convergence occurs with probability at least $1 - \delta$, and then noting that uniform convergence implies $\varepsilon(h)$ is at most 2γ higher than $\varepsilon(h^*) = \min_{h \in \mathcal{H}} \varepsilon(h)$

Corollary. Let $|\mathcal{H}| = k$, and let any δ, γ be fixed. Then for $\varepsilon(\hat{h}) \leq \min_{h \in \mathcal{H}} \varepsilon(h) + 2\gamma$ to hold with probability at least $1 - \delta$, it suffices that

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

The case of infinite $\mathcal{H}$

When dealing with finite $\mathcal{H}$ it was possible to established bounds by counting the hypotheses. However for infinite $\mathcal{H}$ we need some other way of chracterizing the ability of a function class to separate data points. A powerful way is the *VC dimension*, named after its progenitors Vapnik and Chervonenkis. The VC dimension of a space and function class is the largest number of points that can be arbitraily classified. Let's look at two examples.

Example 1: linear separation of points in a two dimensional space As shown in Fig. 1, three points can be arbitraily classified but four cannot. Thus the VC dimension is 3.

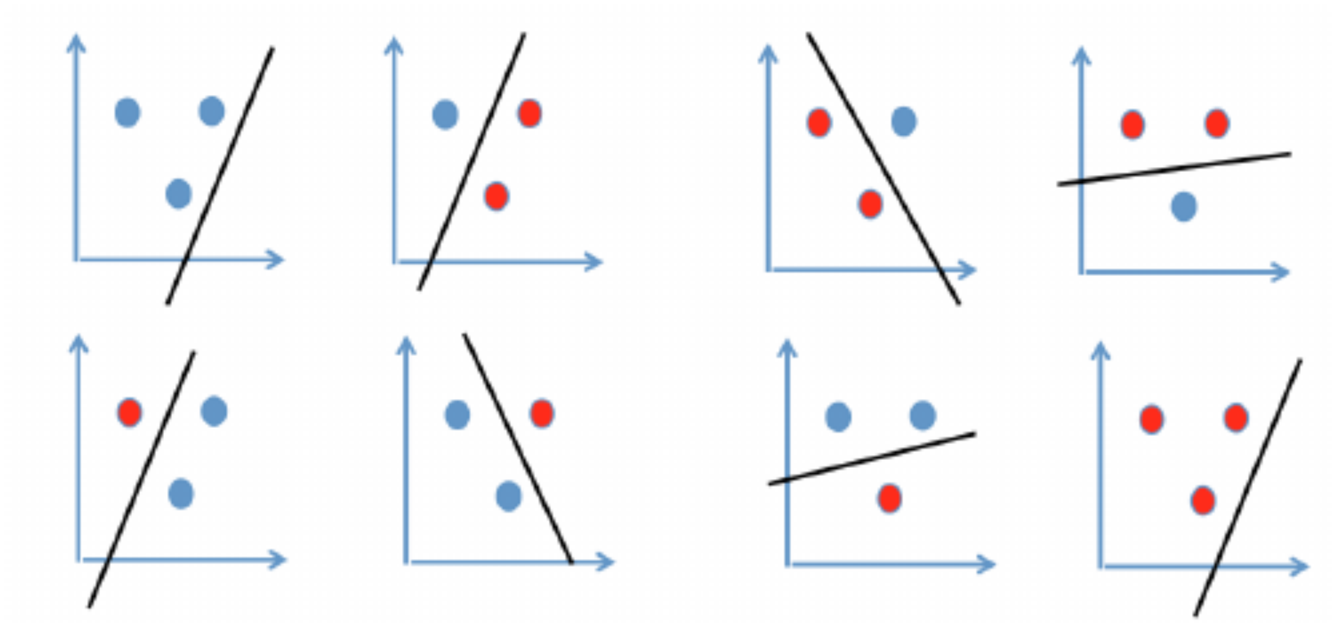
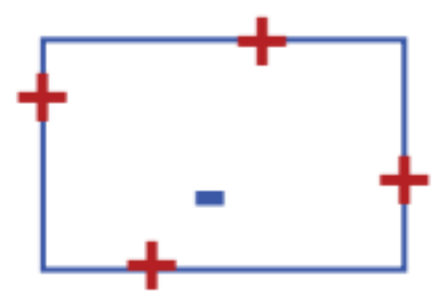


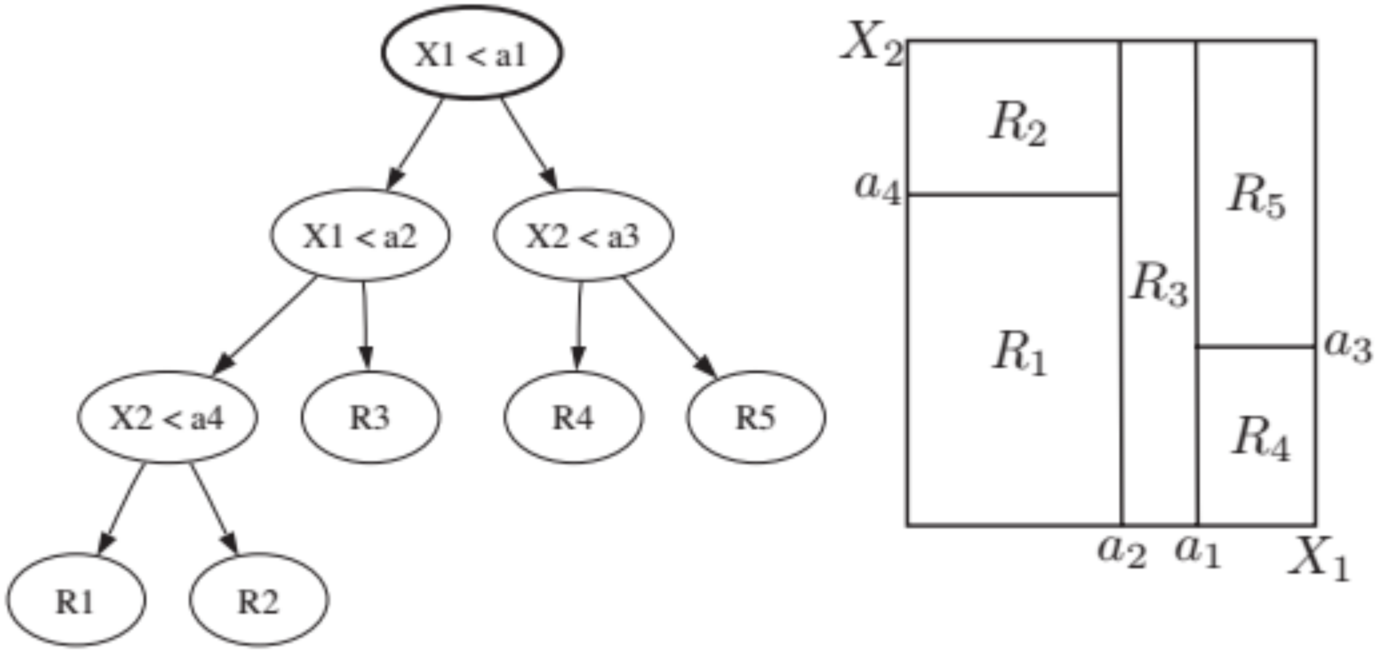
Figure 1: All labelings of three points in two dimensions can be classified by a line.

In three dimensions it turns out that the VC dimension is 4 and in general for n dimensions, the VC dimension is $n + 1$.

Example 2: $\mathcal{H}$ = axis aparallel rectangles.



VC dimension = ?



Presented without proof, the main result

Theorem. Let $\mathcal{H}$ be given, and let $d = \mathrm{VC}(\mathcal{H})$. Then with probability at least $1-\delta$, we have that for all $h \in \mathcal{H}$,

$$|\varepsilon(h) - \hat{\varepsilon}(h)| \le O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right).$$

Thus, with probability at least $1-\delta$, we also have that:

$$\varepsilon(\hat{h}) \le \varepsilon(h^*) + O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right).$$