Statistical Learning Theory

Up to this point we have developed support vector machines purely based on

- linear algebra
- maximum margin

⇒ The key insight was that the maximum margin classifier is the best classifier when considering all possible hyperplanes that separate two classes – we based this argument on optimization theory.

Here we will look at *statistical learning theory* that makes the notion of maximum margin classifier as the optimal classifier rigorous via statistical arguments.

At the heart of this theory is the notion of *VC-dimension*.

VC-Dimension

Informally, the VC-dimension is a measure of the *complexity of a classifier*.

It is a measure of how well the classifiers can separate the points in the input space or model these points without any error.

VC-Dimension

More formally, consider the class of all classifiers with a margin γ of some fixed size, let $\hat{F}[\gamma]$ be that class.

Now consider some dataset D, then the *VC-dimension* of classifiers with the margin γ is the size of the largest subset of points from D that can be separated by classifiers in $\hat{F}[\gamma]$ without any errors for *all possible binary label assignments*.

If all points in D can be separated for all possible label assignments then we say that $\hat{F}[\gamma]$ *shatters* the dataset D.

Definition: The *VC-dimension* of a model class $\hat{F}[\gamma]$ defined over some data set D is the size of the largest finite subset of D shattered by $\hat{F}[\gamma]$.

VC-Dimension

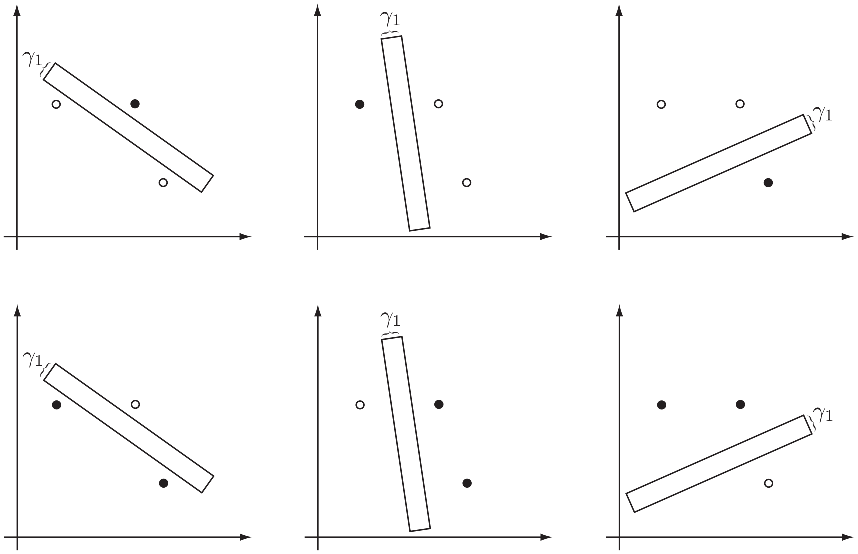
Example: Consider a class of classifiers $\hat{F}[\gamma_1]$ with γ_1 denoting a margin of some fixed size.

Let our dataset be a set of points in two dimensional real space, $D \subset \mathbb{R}^2$.

Let $|D| = 3$, that is D contains three points.

Given a size of γ_1 such that we can separate all three points for all possible label assignments, then the VC-dimension is,

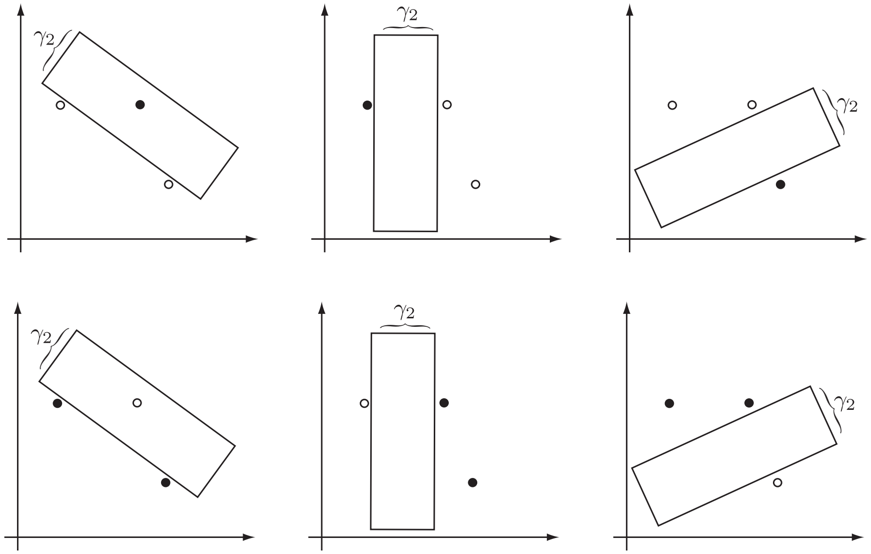
$$h_1 = 3.$$



Now, since $h = |D|$, we say that $\hat{F}[\gamma_1]$ shatters D.

VC-Dimension

Example: Consider a second class of classifiers $\hat{F}[\gamma_2]$ over the same dataset D with $\gamma_2 > \gamma_1$. In particular, the size of γ_2 is such that the classifiers will not be able to separate all points perfectly.



Here we see that the maximum number of points that can be separated by the classifiers in $\hat{F}[\gamma_2]$ is 2. Therefore, we say that the VC-dimension is $h_2 = 2$.

Observe that with $\gamma_2 > \gamma_1$ we have $h_2 < h_1$, that is, models with large margins are less complex that models with small margins.

VC-Dimension

High VC-dimension numbers represent classes of models with high complexity and vice versa,

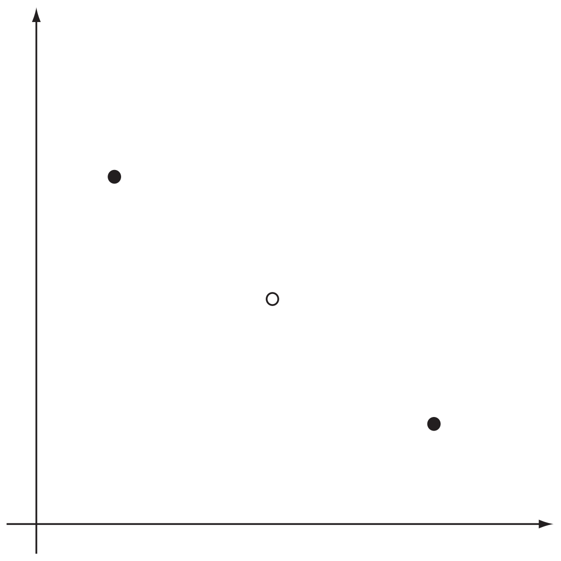
$$\text{Model Complexity} \propto \text{VC-Dimension.}$$

The VC-dimension can be thought of as a formalization of the trade-off between complexity and accuracy of a model. Classifiers with small margins and large VC-dimensions will induce accurate but very complex decision surfaces whereas classifiers with large margins and small VC-dimensions will induce less accurate and less complex decision surfaces.

It turns out that statistical learning theory sheds some light on exactly this trade-off in what is called *expected risk minimization*.

VC-Dimension

Observation: The VC-dimension of a classifier is *data dependent*.



Note, the above dataset VC-dimension $h = 2$ regardless the size of the margin γ in the class of classifiers $\hat{F}[\gamma]$ considered for modeling this dataset.

Mathematical Expectation

$$E[g] = \int_x g(x)P(x)dx,$$

where $g(x)$ is a function over some domain X such that $x \in X$ and $P(x)$ is a probability distribution over X.

$E[g]$ represents the sum of function evaluations over the domain X weighted by their probabilities.

If the domain X is discrete with k elements $x_1, \ldots, x_k$, then the expectation is expressed as,

$$E[g] = \frac{1}{k}\sum_{i=1}^{k} g(x_i).$$

Typically we call the expected value $E[f]$ the *average* or *mean* over all function evaluation over the domain X.

Expected Risk

Assume that $P(\overline{x}, y)$ is the joint probability of the data instances $\overline{x} \in \mathbb{R}^n$ and their corresponding labels $y \in \{+1, -1\}$, also assume that L is the 0-1 loss function, then the *expected loss* for some model $\hat{f} \in \hat{F}[\gamma]$ defined over the data universe is,

$$E[L(y, \hat{f}(\overline{x}))] = \int L(y, \hat{f}(\overline{x}))\, dP(\overline{x}, y).$$

In other words, the expected loss is the expected number of mistakes a model will commit over the underlying data universe.

We often write

$$R[\hat{f}] = E[L(y, \hat{f}(\overline{x}))],$$

where $R[\hat{f}]$ is called the *expected risk*.

Risk Minimization

The goal in statistical learning is to find a model $\hat{f}^* \in \hat{F}$ that minimizes the expected risk

$$\hat{f}^* = \operatorname*{argmin}_{\hat{f} \in \hat{F}} R[\hat{f}],$$

where $\hat{F}$ represents the class of all model classes such that $\hat{F}[\gamma] \subset \hat{F}$ for all margins γ.

Unfortunately, this is impossible in the present formulation of the expected risk because we do not know the probability distribution $P(\overline{x}, y)$.

If we did, there would be nothing to learn.

Empirical Risk

However, we do have some information on the joint probability distribution in the form of samples in our dataset D,

$$D = \{(\overline{x}_1, y_1), \ldots, (\overline{x}_l, y_l)\} \subset \mathbb{R}^n \times \{+1, -1\}.$$

We can use these samples to estimate the risk, we call this the *empirical risk* $R_{\text{emp}}[\hat{f}]$ of some model $f \in F$ and define it as

$$R_{\text{emp}}[\hat{f}] = E[L(y, \hat{f}(\overline{x}))] = \frac{1}{l}\sum_{i=1}^{l} L(y_i, \hat{f}(\overline{x}_i)),$$

where $(\overline{x}_i, y_i) \in D$. Then,

$$\begin{aligned} \hat{f}^* &= \operatorname*{argmin}_{\hat{f}\in\hat{F}} R_{\text{emp}}[\hat{f}] \\ &= \operatorname*{argmin}_{\hat{f}\in\hat{F}} \left(\frac{1}{l}\sum_{i=1}^{l} L(y_i, \hat{f}(\overline{x}_i)) \right). \end{aligned}$$

However, minimizing this equation is overly optimistic, since we can always find some model which fits the sample data extremely well.[a]

[a] Compare this to the training error of a model.

VC-Confidence

In order to use the empirical risk for estimating the best model for the expected risk, Vapnik introduced a new term called the *VC-confidence* which together with the empirical risk can be considered a bound on the expected risk,

$$R[\hat{f}] \leq \underbrace{R_{\text{emp}}[\hat{f}] + \overbrace{v(l,h,\eta)}^{\text{VC-confidence}}}_{\text{VC generalization bound}},$$

where h is the VC-dimension of $\hat{f}$ and η is some small number such that $0 < \eta < 1$. Typically $\eta = 0.05$ for the 95% confidence interval, since the theory states that the bound holds with probability $1 - \eta$.

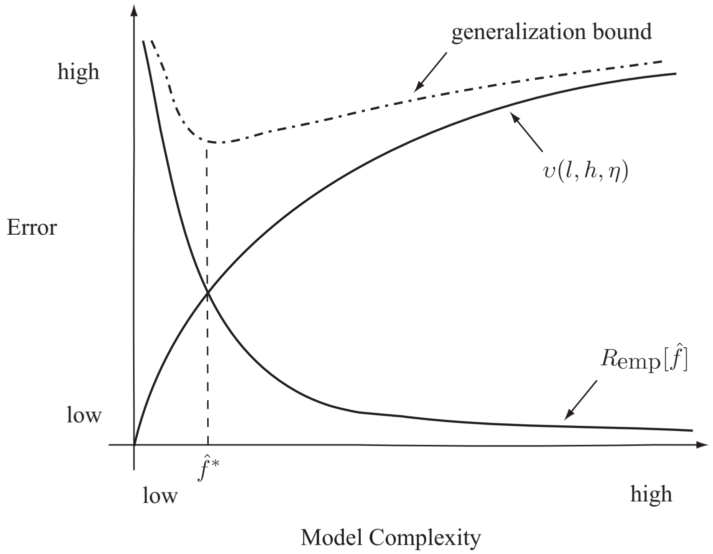
The VC-confidence term is defined as follows,

$$v(l,h,\eta) = \sqrt{\frac{h(log(\frac{2l}{h}) + 1) - log(\frac{\eta}{4})}{l}}.$$

Notice that the right side of the above equations do not depend the joint probability distribution $P(\overline{x}, y)$.

Observation: v is directly proportional to h and indirectly proportional to the number of training instances l.

Generalization Bound



The figure illustrates the relationship between the empirical risk $R_{\text{emp}}[\hat{f}]$ and the VC-confidence $v(l, h, \eta)$.

Note that as the complexity of the models increases the empirical risk decreases. That is, complex models allow us to model the training data well.

On the other hand, as model complexity increases so does the VC-confidence. Here, complex models will commit more errors on data not contained in the training data.

The generalization bound can be considered the envelope of these two curves. It is interesting to note that minimizing the generalization bound is equivalent to making just the right trade-off between model complexity and error rate $\Rightarrow \hat{f}^*$.

Structural Risk Minimization

With the generalization bound we now have a way to characterize models that optimally trade off complexity and error.

The question remains, how do we find these models?

Our notion of model class $\hat{F}$ is most likely infinite and we have to traverse this class of models to find the optimal model $\hat{f}^*$ that minimizes the generalization bound.

An effective way to traverse this model class in search for an optimal model is *structural risk minimization*.

Structural Risk Minimization

suppose we have a class of linear models $\hat{F}$ with

$$\hat{F}[\gamma_1], \ldots, \hat{F}[\gamma_k] \subset \hat{F},$$

where

$$\hat{F}[\gamma_1] \subset \hat{F}[\gamma_2] \subset \ldots \subset \hat{F}[\gamma_k] \text{ if } h_1 < h_2 < \ldots < h_k,$$

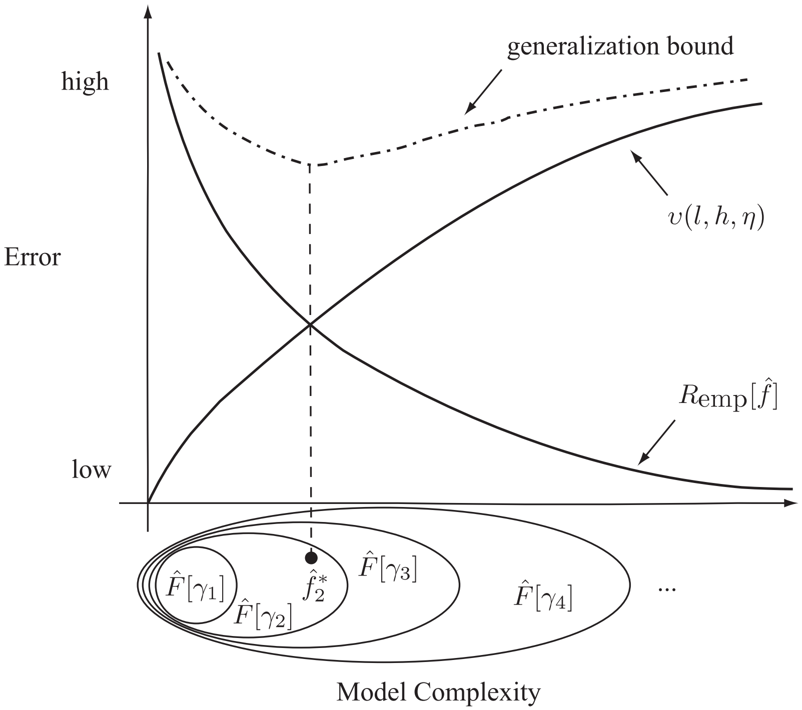
where h_i is the VC-dimension of model class $\hat{F}[\gamma_i]$.

Given that we assume linear models, the equation above implies that the margins of the various model classes are also partially ordered,

$$\gamma_k < \ldots < \gamma_2 < \gamma_1.$$

This gives us an effective procedure to find the optimal model: We start with the least complex model class $\hat{F}[\gamma_1]$ and minimize the generalization bound. We then move on to the next model class, in this case $\hat{F}[\gamma_2]$, and compute the optimal model $\hat{f}^*_2$ in a similar fashion. We terminate our search if we find that the generalization bound of some model $\hat{f}^*_{i+1} \in \hat{F}[\gamma_{i+1}]$ is larger than the generalization bound of the model $\hat{f}^*_i \in \hat{F}[\gamma_i]$. In this case, $\hat{f}^*_i$ is the optimal model.

Structural Risk Minimization



Here we see that statistical learning theory and our intuitive notion of maximum margin classifier coincide.

In addition, statistical learning theory provides a nice mathematical framework for our intuitions.