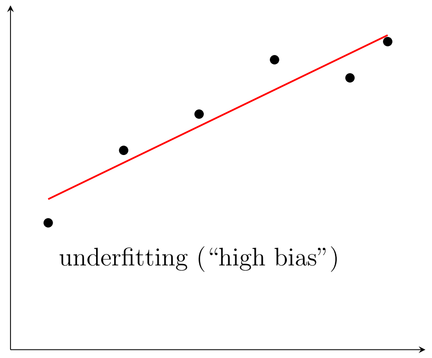
Lecture 8

1. Bias / variance
2. Empirical risk minimization (ERM)
3. Union bound / Hoeffding inequality
4. Uniform convergence
5. VC dimension

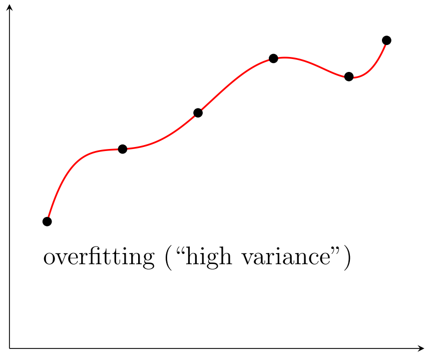
1 Bias / variance

In most cases to apply supervised machine learning algorithms we can use some packages where these algorithms are already implemented. But even when we use built-in algorithms we should know how it works. In this lecture we try to obtain some theoretical results about machine learning algorithms in general (without concentration on some particular algorithm).

In the Lecture 2 we talked about overfitting and underfitting. Let us recall these notions:

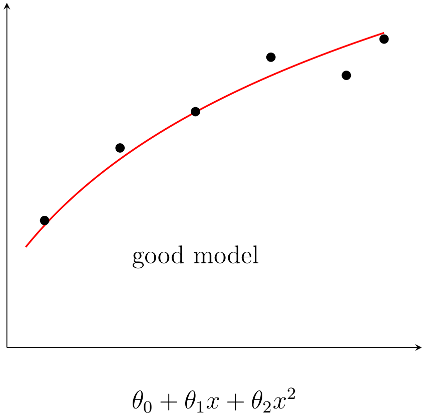


$\theta_0 + \theta_1 x$

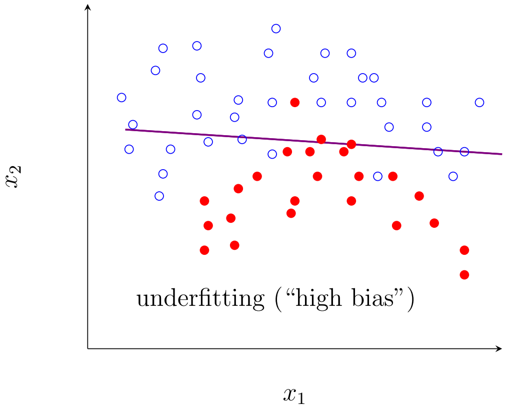


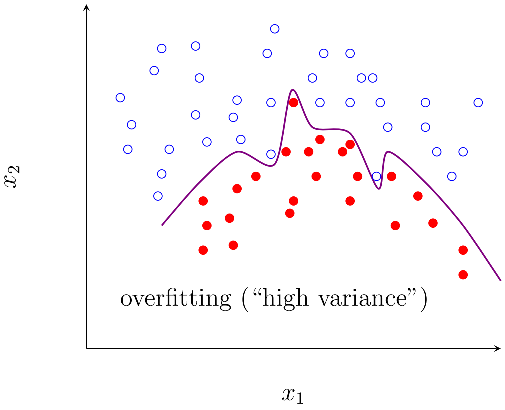
$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

On the first picture we see that even when we take more data our error does not decrease, we say that this model has "high bias", or underfitting. On the second picture the error is zero, but if we add more data the prediction will be very bad, we say that this model has "high variance", or overfitting. The best model lies somewhere between, for example, it could be some quadratic polynomial in our example:

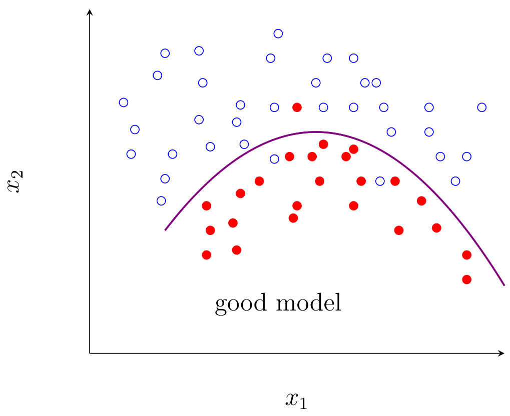


The similar situation is happening for classification problems, for example, the next two figures show the underfitting and overfitting models for 2-class classification problem:





On the first picture we fit logistic regression with linear function (underfitting, or "high bias"), on the second we fit logistic regression with higher order polynomial (overfitting, or "high variance"). Our purpose is to find the best model somewhere between, for example, as before the good model could be the logistic regression with quadratic polynomial:



2 Empirical risk minimization (ERM)

To formalize the problem, consider a simple linear classification problem with the hypothesis:

$$h_\theta(x) = g(\theta^T x),$$

where

$$g(z) = \mathbb{1}\{z \geqslant 0\}$$

and $y \in \{0,1\}$. Denote the training dataset by $S = \{x^{(i)}, y^{(i)}\}_{i=1}^m$ and assume that $(x^{(i)}, y^{(i)})$ are independent identically distributed sampled from some distribution D.

The **training error**, or **risk**, for this problem is defined as

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

The **empirical risk minimization (ERM)** problem is a non-convex optimization problem that can be formulated as

$$\hat{\theta} = \arg\min_\theta \hat{\varepsilon}_S(h_\theta).$$

With this formulation we can define the parametric models only. To cover nonparametric models we formulate the problem in more general way. Let the hypothesis class $\mathcal{H} = \{h_\theta, \theta \in \mathbb{R}^{n+1}\}$ includes the hypotheses $h_\theta : X \to \{0,1\}$. The hypothesis class $\mathcal{H}$ can contain any set of functions, for example, functions, produced by neural networks, or piecewise functions, produced by tree-based methods. Then ERM problem is equivalent to

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\varepsilon}_S(h).$$

The **generalization error** for the hypothesis h is defined as

$$\varepsilon(h) = P(h(x) \neq y \,|\, (x, y) \sim D) \tag{2}$$

(probability that hypothesis h does not classify training example x correctly, given that (x, y) are drawn from the distribution D). The main difference between training error and generalization error is that **training error is calculated on the training set** S only, but **generalization error is calculated for any sample drawn from the distribution** D. The problem of generalization error minimization is very complicated and cannot be solved in majority of cases. Usually, this problem is replaced by the ERM problem. The main question what is the relationship between training error $\hat{\varepsilon}$ and generalization error ε.

3 Union bound / Hoeffding inequality

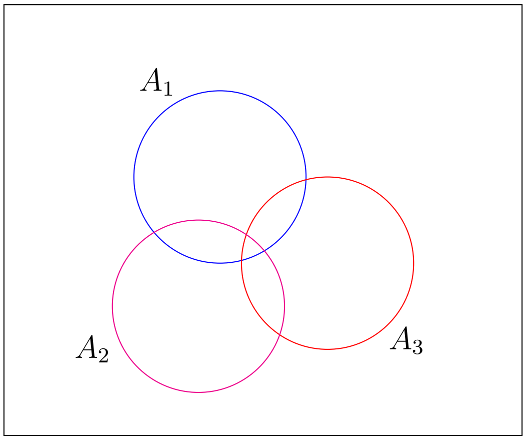
Two facts from the probability theory are necessary for our next results.

Theorem (union bound). Let $A_1, A_2, \ldots, A_k$ be k probabilistic events (not necessarily independent). Then

$$P(A_1 \cup A_2 \cup \ldots \cup A_k) \leqslant P(A_1) + P(A_2) + \ldots + P(A_k),$$

where $\cup$ defines the union of events $A_1, \ldots, A_k$ (operator "OR").

The illustration for the union bound theorem is the Venn diagram, for example, for 3 events it looks as



Theorem (Hoeffding inequality). Let $z_1, \dots, z_m$ be m independent identically distributed Bernoulli random variables with mean φ, i.e.

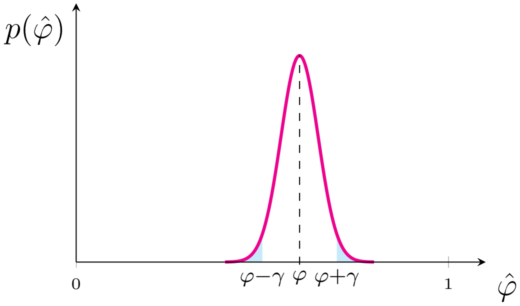
$$P(z_i = 1) = \varphi \text{ for all } i,$$

and $\hat{\varphi} = \frac{1}{m}\sum_{i=1}^{m} z_i$, then

$$P(|\hat{\varphi} - \varphi| > \gamma) \leqslant 2e^{-2\gamma^2 m}$$

for any fixed number $\gamma > 0$.

This theorem means that if we choose a number φ, then the density function for $\hat{\varphi}$ has a shape of normal distribution (the distribution of $\hat{\varphi}$ is roughly normal):



When γ is fixed then the probability to have $\hat{\varphi}$ on the tails (shaded regions) is less than $2e^{-2\gamma^2 m}$, and it is decreasing exponentially with m.

4 Uniform convergence

Consider the case when the hypothesis class $\mathcal{H}$ is finite: $\mathcal{H} = \{h_1, \ldots, h_k\}$ (we have just k hypotheses in the hypothesis class $\mathcal{H}$). The solution for the ERM problem is straightforward: for each hypothesis $h_j \in \mathcal{H}$ calculate the training error $\hat{\varepsilon}$ on the training set S and choose the hypothesis with the minimal error:

$$\hat{h} = \arg\min_{h_j \in \mathcal{H}} \hat{\varepsilon}_S(h_j).$$

We prove that such simple procedure helps to minimize generalization error ε as well. For any $h_j \in \mathcal{H}$, we define

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

By definition (2),

$$P(z_i = 1) = \varepsilon(h_j),$$

which means that z_i are independent identically distributed Bernoulli random variables with mean $\varepsilon(h_j)$. The formula (1) implies

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

then by Hoeffding inequality

$$P(|\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma) \leqslant 2e^{-2\gamma^2 m}.$$

The obvious consequence we can obtain from this inequality: if m becomes large then the difference between training error and generalization error becomes smaller.

For any hypothesis $h_j \in \mathcal{H}$ we define the random event A_j that occurs if $|\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma$. We proved that

$$P(A_j) \leqslant 2e^{-2\gamma^2 m}.$$

Consider the probability

$$\begin{aligned} &P(\exists h_j \in \mathcal{H} \text{ such that } |\varepsilon(h_j) - \hat{\varepsilon}(h_j)| > \gamma) = \\ &= P(A_1 \cup A_2 \cup \ldots \cup A_k) \leqslant \sum_{j=1}^{k} P(A_j) = \sum_{j=1}^{k} 2e^{-2\gamma^2 m} = 2ke^{-2\gamma^2 m} \end{aligned}$$

by the union bound theorem. The probability of opposite event

$$P(\forall h_j \in \mathcal{H} : |\varepsilon(h_j) - \hat{\varepsilon}(h_j)| \leqslant \gamma) \geqslant 1 - 2ke^{-2\gamma^2 m}.$$

The last result can be reformulated as: with probability $1 - 2ke^{-2\gamma^2 m}$ the training error $\hat{\varepsilon}(h)$ will be within γ of generalization error $\varepsilon(h)$ for all $h \in \mathcal{H}$. This statement is called **uniform convergence**.

There are three parameters of interest:

- γ: how far the training error from the generalization error (**error bound**)
- m: how many training examples we have
- δ: what is the range for probability

We analyse the uniform convergence result:

- Given γ and m, the probability $\delta = 2ke^{-2\gamma^2 m}$.
- Given γ and δ, what should be the size m of the training set?

  $$\delta = 2ke^{-2\gamma^2 m} \Leftrightarrow m = \frac{1}{2\gamma^2} \ln \frac{2k}{\delta},$$

  which means that if the number of training examples $m \geqslant \frac{1}{2\gamma^2} \ln \frac{2k}{\delta}$, then with probability $(1 - \delta)$ we have $|\varepsilon(h) - \hat{\varepsilon}(h)| \leqslant \gamma$ for all $h \in \mathcal{H}$ (**"sample complexity" bound**).
- Given m and δ, what will be the error bound γ?

  $$\delta = 2ke^{-2\gamma^2 m} \Leftrightarrow \gamma = \sqrt{\frac{1}{2m} \ln \frac{2k}{\delta}},$$

  which means that with probability $(1 - \delta)$

  $$|\varepsilon(h) - \hat{\varepsilon}(h)| \leqslant \sqrt{\frac{1}{2m} \ln \frac{2k}{\delta}}$$

  for any $h \in \mathcal{H}$.

Notice that m grows as $\ln k$ and if we add huge number of hypothesis to $\mathcal{H}$, then m does not grow a lot.

Let us see how the uniform convergence result is related to ERM problem. We assume

$$|\varepsilon(h) - \hat{\varepsilon}(h)| < \gamma \tag{3}$$

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \hat{\varepsilon}(h) \text{ (ERM problem)} \tag{4}$$

$$h^* = \arg\min_{h\in\mathcal{H}} \varepsilon(h) \text{ (best hypothesis)} \tag{5}$$

The equation (5) defines the best hypothesis for the specific problem. As we have limited resources usually we cannot find it, but we can approximate it by the hypothesis $\hat{h}$.

Theorem. Let $|\mathcal{H}| = k$ and let m, δ be fixed, then with probability $1-\delta$

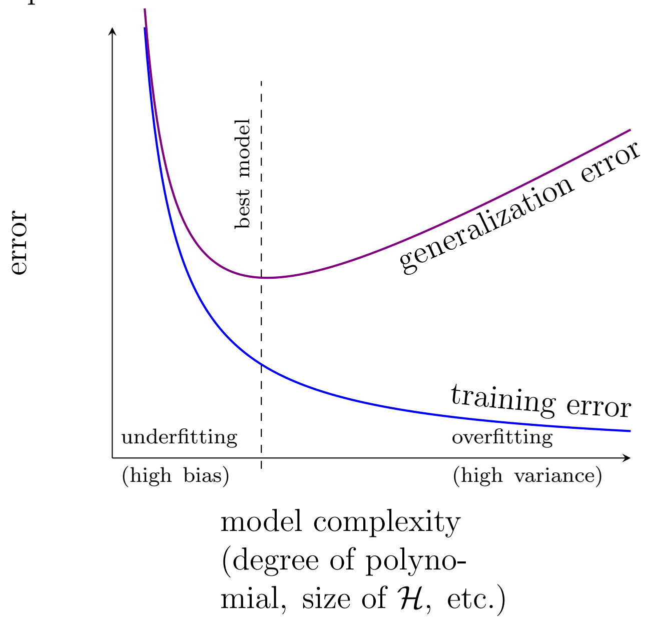
$$\varepsilon(\hat{h}) \leqslant \left(\min_{h\in\mathcal{H}} \varepsilon(h)\right) + 2\sqrt{\frac{1}{2m}\ln\frac{2k}{\delta}}. \tag{6}$$

Proof. By equation (3) and (4)

$$\varepsilon(\hat{h}) \overset{(3)}{\leqslant} \hat{\varepsilon}(\hat{h}) + \gamma \overset{(4)}{\leqslant} \hat{\varepsilon}(h^*) + \gamma \overset{(3)}{\leqslant} \varepsilon(h^*) + 2\gamma \overset{(5)}{=} \left(\min_{h\in\mathcal{H}} \varepsilon(h)\right) + 2\gamma.$$

Before we proved that $\gamma = \sqrt{\frac{1}{2m}\ln\frac{2k}{\delta}}$ and the equation (3) holds with probability $(1-\delta)$ for any hypothesis $h \in \mathcal{H}$, that proves the theorem. □

How we can use the obtained result? We could start solving the machine learning problem in the class of linear hypothesis $\mathcal{H}$ and after expand the class $\mathcal{H}$ to the class $\mathcal{H}'$ of quadratic hypothesis. What happens with our errors? Obviously, that $\varepsilon(h^*)$ becomes better (the best hypothesis on the class of quadratic functions is better then on the class of linear functions). But simultaneously, the number of hypothesis k increases. If we say that the first term $\min\limits_{h\in\mathcal{H}} \varepsilon(h)$ in the equation (6) represents the "bias" and the second term $2\sqrt{\frac{1}{2m}\ln\frac{2k}{\delta}}$ represent the "variance", then if we expand the hypothesis class $\mathcal{H}$ our "bias" term is decreasing and "variance" term is increasing. This behaviour can be visualized with the graph for fixed number m of training examples:



Corollary. Let $|\mathcal{H}| = k$ and let δ and γ be fixed. Then to have

$$\varepsilon(\hat{h}) \leqslant \min_{h \in \mathcal{H}} \varepsilon(h) + 2\gamma$$

with probability $(1 - \delta)$, it is sufficient that

$$m \geqslant \frac{1}{2\gamma^2} \ln \frac{2k}{\delta} = O\left(\frac{1}{\gamma^2} \ln \frac{k}{\delta}\right).$$

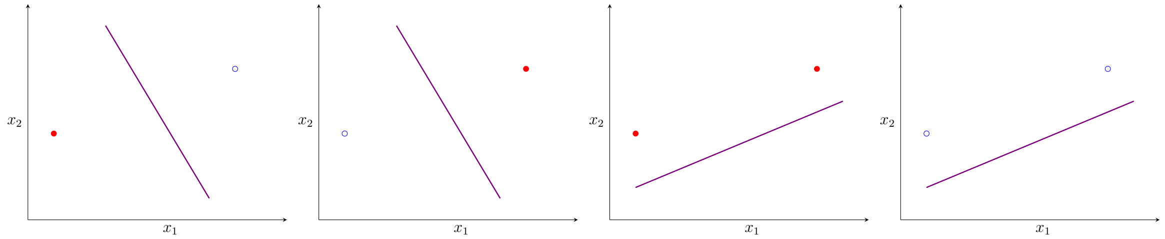
Example. If $k = 100$ (number of hypotheses), $\gamma = 0.1$ (the desired difference between training error and generalization errors), $\delta = 0.8$ (the probability), then $m = \dfrac{1}{0.1^2} \ln \dfrac{100}{0.8} \approx 483$.

5 VC dimension

Definition. Given a set of points $S = \{x^{(1)}, \ldots, x^{(m)}\}$, we say a hypothesis class $\mathcal{H}$ **shatters** S if $\mathcal{H}$ can realize any labelling on it (for any set of labels for m points there exists the hypothesis $h \in \mathcal{H}$ such that it maps points perfectly on the chosen set of labels).

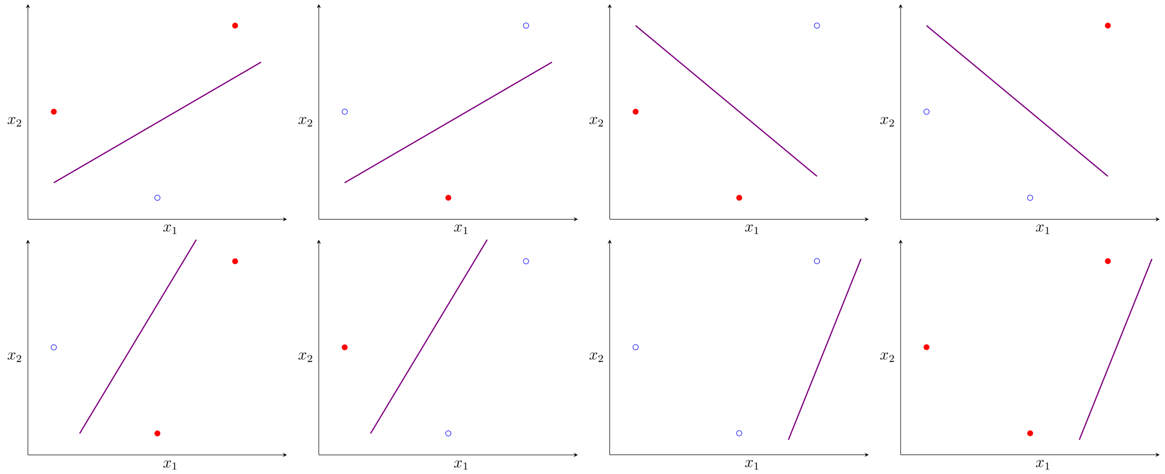
Examples.

1. $\mathcal{H}$ is a class of linear classifiers in 2D and $S = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)})\}$, then all possible labelings are



For all labelings we can find the linear classifier which means that $\mathcal{H}$ shatters S.

2. $\mathcal{H}$ is a class of linear classifiers in 2D and $S = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)})\}$, then all possible labelings are

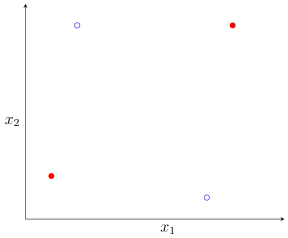


As before for all labelings we can find the linear classifier which means that $\mathcal{H}$ shatters S.

3. $\mathcal{H}$ is a class of linear classifiers in 2D and

$$S = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), (x^{(4)}, y^{(4)})\},$$

then we can find labelling for which we cannot find the linear classifier that perfectly splits our classes:

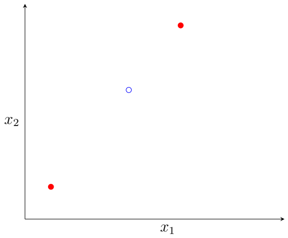


In this situation we say that $\mathcal{H}$ does not shatter data S.

Definition. The **Vapnik-Chervonenkis dimension** $VC(\mathcal{H})$ of the hypothesis class $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

Example. From the previous examples we can say that if $\mathcal{H}$ is a set of linear classifiers in 2D, then $VC(\mathcal{H}) = 3$.

Notice that there is a set of 3 points on the plane that cannot be shattered by linear classifiers $\mathcal{H}$:



but it is enough to say that there exist 3 points such that it is shattered by $\mathcal{H}$. For 4 points we can choose any configuration and for any configuration it cannot be shattered by $\mathcal{H}$.

More generally, in the n-dimensional space

$$VC(\{\text{linear classifiers in } \mathbb{R}^n\}) = n + 1.$$

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

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

Thus, with probability $(1 - \delta)$ we also have

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

Corollary. In order to guarantee that

$$\varepsilon(\hat{h}) \leqslant \varepsilon(h^*) + 2\gamma$$

with probability $(1-\delta)$ it suffices that

$$m = O_{\gamma,\delta}(d).$$

(the last means that if we fix γ and δ, then $m = O(d)$).

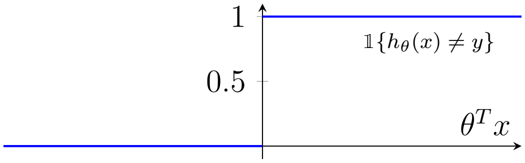
In many cases VC dimension is very close to the number of parameters, but there is one interesting case when it is not true. For the large margin classifiers (for example, SVM) the following property holds.

Theorem. Let $\mathcal{H}$ be the hypothesis class of the large margin classifiers and $||x||_2 \leqslant R$, then

$$VC(\mathcal{H}) \leqslant \left\lceil \frac{R^2}{4\gamma^2} \right\rceil + 1,$$

where γ is a minimal distance to the decision boundary.

The last remark is that in this lecture we considered step function as our risk function for the ERM problem.



It turns out that logistic regression and SVM are just approximations to the ERM problem with logistic loss and hinge loss accordingly (as step function is not convex function). And this fact means that all the theory we have developed is correct for logistic regression and SVM classifiers.