Empirical Risk Minimization

Learner outputs $h_S : \mathcal{X} \to \mathcal{Y}$.

from the training set

Goal: find h_S which minimizes the generalization error $L_{\mathcal{D},f}(h)$

$L_{\mathcal{D},f}(h)$ is unknown!

$\frac{\# \text{ wrong}}{\# \text{ instances}}$

What about considering the error on the training data, that is, reporting in output h_S that minimizes the error on training data?

It's a function of the hypothesis

m = # istances in the training set

Training error: $L_S(h) \stackrel{def}{=} \frac{|\{i : h(x_i) \neq y_i, 1 \leq i \leq m\}|}{m}$

# of instances $\in S$ for which h predicts the wrong label

Note: the *training error* is also called *empirical error or empirical risk*

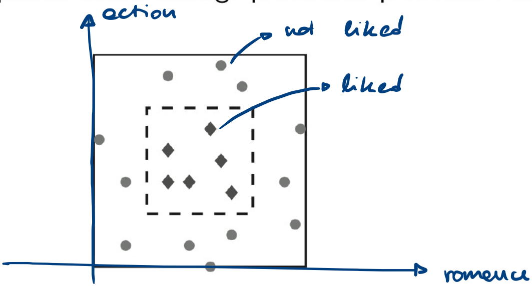
smallest training error

Empirical Risk Minimization (ERM): produce in output h minimizing $L_S(h)$

we assume there's a link between the training set and the "future data" (same probability distribution)

What can go wrong with ERM?

Consider our simplified movie ratings prediction problem. Assume data is given by:



Assume $\mathcal{D}$ and f are such that:

- instance x is taken uniformly at random in the square ($\mathcal{D}$)
- label is 1 if x inside the inner square, 0 otherwise (f)
- area inner square $= 1$, area larger square $= 2$

Consider classifier given by

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in \{1, \ldots, m\} : x_i = x \\ 0 & \text{otherwise} \end{cases}$$

if x is in the training set

Is it a good predictor?

$L_S(h_S) = 0$ but $L_{\mathcal{D},f}(h_S) = 1/2$ → whenever x is in the inner square (and was not in the training set)

Good results on training data but poor generalization error

$\Rightarrow$ **overfitting**

When does ERM lead to good performances in terms of generalization error?

→ - Sufficient data
- Robust feature representation
- Similarity between train and test data