Goal of Learning Algorithms

- The early learning algorithms were designed to find such an accurate fit to the data.
- A classifier is said to be *consistent* if it performed the correct classification of the training data
- The ability of a classifier to correctly classify data *not in the training set* is known as its *generalization*
- Bible code? 1994 Taipei Mayor election?
- Predict the real future *NOT fitting the data in your hand or predict the desired results*

Probably Approximately Correct Learning
pac Model

- Key assumption:

  Training and testing data are generated *i.i.d.* according to an *fixed but unknown* distribution $\mathcal{D}$

- When we evaluate the "quality" of a hypothesis (classification function) $h \in H$ we should take the *unknown* distribution $\mathcal{D}$ into account (*i.e.* "average error" or "expected error" made by the $h \in H$)

- We call such measure risk functional and denote it as $\underset{\mathcal{D}}{err}(h) = \mathcal{D}\{(x, y) \in X \times \{1, -1\} \mid h(x) \neq y\}$

Generalization Error of pac Model

- Let $S = \{(x^1, y_1), \ldots, (x^l, y_l)\}$ be a set of l training examples chosen i.i.d. according to $\mathcal{D}$
- Treat the generalization error $\underset{\mathcal{D}}{err}(h_S)$ as a *r.v.* depending on the random selection of S
- Find a bound of the trail of the distribution of *r.v.* $\underset{\mathcal{D}}{err}(h_S)$ in the form $\varepsilon = \varepsilon(l, H, \delta)$
- $\varepsilon = \varepsilon(l, H, \delta)$ is a function of l, H and δ,where $1 - \delta$ is the confidence level of the error bound which is given by learner

Probably Approximately Correct

- We assert:

$$Pr(\{\, err_{\mathcal{D}}(h_S) > \varepsilon = \varepsilon(l, H, \delta)\}) < \delta$$

or

$$Pr(\{\, err_{\mathcal{D}}(h_S) \leqslant \varepsilon = \varepsilon(l, H, \delta)\}) \geqslant 1 - \delta$$

- The error made by the hypothesis h_s will be less then the error bound $\varepsilon(l, H, \delta)$ that is not depend on the unknown distribution $\mathcal{D}$

Find the Hypothesis with Minimum Expected Risk?

- Let $S = \{(x^1, y_1), \ldots, (x^l, y_l)\} \subseteq X \times \{-1, 1\}$ be the training examples chosen i.i.d. according to $\mathcal{D}$ with the probability density $p(x, y)$
- The expected misclassification error made by $h \in H$ is

$$R[h] = \int_{X\times\{-1,1\}} \frac{1}{2}|h(x) - y|dp(x, y)$$

- The *ideal* hypothesis h^*_{opt} should has the smallest expected risk $R[h^*_{opt}] \leqslant R[h], \quad \forall h \in H$

Unrealistic !!!

Empirical Risk Minimization (ERM)

($\mathfrak{D}$ and $p(x,y)$ are not needed)

- Replace the expected risk over $p(x,y)$ by an average over the training example
- The empirical risk: $R_{emp}[h] = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{2}|h(x^i) - y_i|$
- Find the hypothesis h^*_{emp} with the smallest empirical risk $R_{emp}[h^*_{emp}] \leqslant R_{emp}[h], \quad \forall h \in H$
- Only focusing on empirical risk will cause *overfitting*

VC Confidence

(The Bound between $R_{emp}[h]$ & $R[h]$)

- The following inequality will be held with probability $1-\delta$

$$R[h] \leqslant R_{emp}[h] + \sqrt{\frac{v(log(2l/v)+1)-log(\delta/4)}{l}}$$

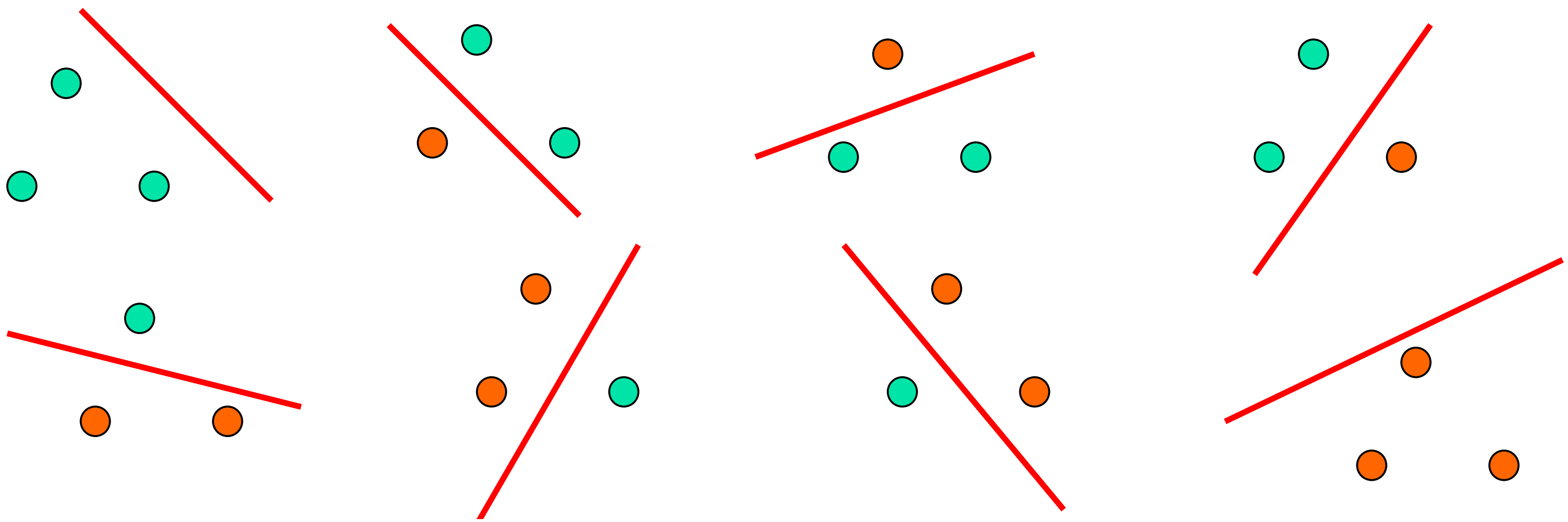
C. J. C. Burges, *A tutorial on support vector machines for pattern recognition*,
Data Mining and Knowledge Discovery 2 (2) (1998), p.121-167

Why We Maximize the Margin? (Based on Statistical Learning Theory)

- The Structural Risk Minimization (SRM):
  - The expected risk will be less than or equal to empirical risk (training error)+ VC (error) bound
- $\|w\|_2 \propto VC \;\; bound$
- $\min VC \;\; bound \Leftrightarrow \min \frac{1}{2}\|w\|_2^2 \Leftrightarrow \max Margin$

Capacity (Complexity) of Hypothesis Space H :VC-dimension

- A given training set S is *shattered* by H if and only if for every labeling of S, $\exists\ h \in H$ consistent with this labeling
- Three (linear independent) points shattered by a hyperplanes in R^2



Shattering Points with Hyperplanes in R^n

Can you always shatter three points with a line in R^2?

Theorem: Consider some set of m points in R^n. Choose a point as origin. Then the m points can be shattered by oriented hyperplanes if and only if the position vectors of the rest points are linearly independent.

Definition of VC-dimension

(A Capacity Measure of Hypothesis Space H)

- The *Vapnik-Chervonenkis* dimension, $VC(H)$, of hypothesis space H defined over the input space X is the size of the (existent) largest finite subset of X shattered by H
- If arbitrary large finite set of X can be shattered by H, then $VC(H) \equiv \infty$
- Let $H = \{all\ \ hyperplanes\ \ in\ \ R^n\}$ then $VC(H) = n + 1$