PAC Learning

Ludwig Krippahl

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

PAC Learning

Summary

- Empirical Risk Minimization
- Probably Approximately Correct Learning
- Shattering
- VC Dimension

- Previously, we saw Bias-Variance tradeoff
  - High bias, underfitting; high variance, overfitting
  - How to select? Empirically (cross-validation)

Today:

- Understand these problems more formally

Empirical Risk Minimization

Empirical Risk Minimization

- Loss: how bad our predictions are
  - Quadratic error, Brier score, 1-Accuracy, ...
- Risk: the expected (average) loss
- Empirical Risk: the measured average loss
- Empirical Risk Minimization
  - Minimize the average loss on the training set

- True risk: average loss over all data
- Empirical risk underestimates true risk (true error)

ERM

Empirical Risk and True Risk

- Union bound: $A_1, A_2, \ldots, A_k$ are random events

$$P(A_1 \cup A_2 \cup \ldots A_k) \leq P(A_1) + P(A_2) + \ldots + P(A_k)$$

- Hoeffding's inequality: if $B_1, B_2, \ldots, B_m$ are i.i.d. Bernoulli(ϕ)

$$P(B_i = 1) = \phi \qquad \hat{\phi} = \frac{1}{m} \sum_{i=1}^{m} B_i$$

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

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

- The probability of average over m {0,1} events deviating γ from the true probability ϕ decreases with m

Empirical Risk Minimization

- Consider binary classifiers, $h : \mathcal{X} \rightarrow \{0, 1\}$
- Given S with m examples from $\mathcal{X}$ with dist. D, the empirical error (training error) is:

$$\hat{E}_S(h) = \frac{1}{m} \sum_{i=1}^{m} 1\{h(x^i) \neq c(x^i)\}$$

- The true error is:

$$E(h) = P_{x \sim D}\left(h(x) \neq c(x)\right)$$

ERM

Empirical Risk Minimization

- Suppose binary classifier with parameters θ
- Best parameters can be found by:

$$\hat{\theta} = \arg\min_{\theta} \hat{E}(h_\theta)$$

- This is empirical risk minimization, which is NP-Hard in general but can be approximated
- And can bound the true error with Hoeffding's inequality

PAC Learning

PAC Learning

Definitions

- $\mathcal{X}$: set of possible examples (instances)
- $c : \mathcal{X} \rightarrow \{0, 1\}$: target function to learn
- $\mathcal{H}$: hypothesis class learner considers
- $\mathcal{D}$: distribution of examples over $\mathcal{X}$
- S: training sample

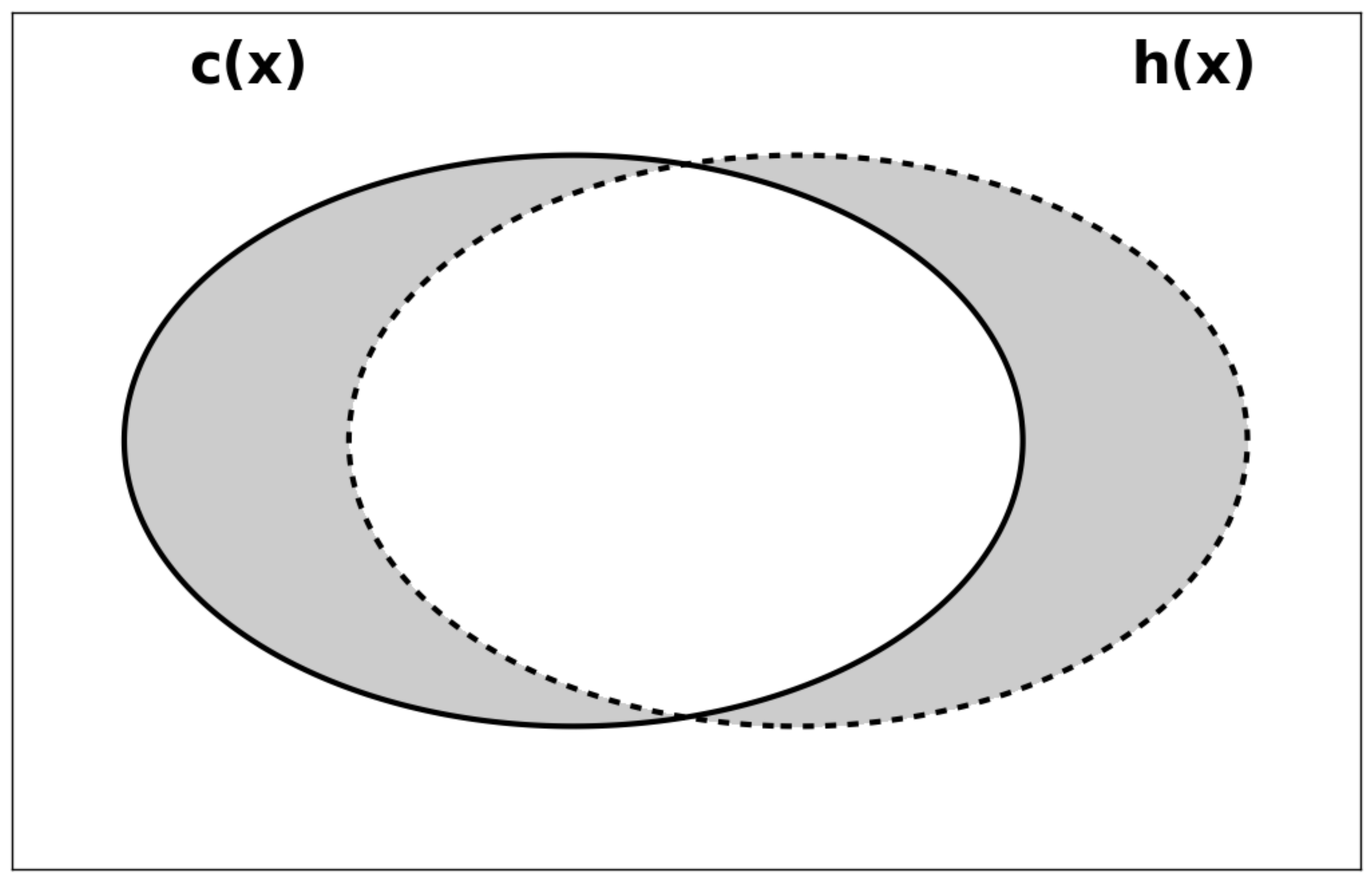
Learning

- Learner receives S from $\mathcal{X}$ with dist. $\mathcal{D}$
- Selects $\hat{h}$ from $\mathcal{H}$ minimizing the empirical error:

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

PAC Learning

- True error of h is $E(h) = P_{x \sim D}\left(h(x) \neq c(x)\right)$



PAC Learning

True error

- True error of h is

$$E(h) = P_{x \sim D}\left(h(x) \neq c(x)\right)$$

- The true error is not directly observable
- Learner can only measure the training error

$$\hat{E}_S(h) = \frac{1}{m} \sum_{i=1}^{m} 1\{h(x^{(i)}) \neq c(x^{(i)})\}$$

- We cannot reasonably demand zero true error

- Not all possible examples in training, so multiple hypotheses seem correct
- Examples may be misleading in their correlation to the classes.

PAC Learning

Probably Approximately Correct Learning

- Weaker requirements:

- Approximately correct: $E(\hat{h}) \leq \epsilon$
- Probably Approximately Correct:

$$P\left(E(\hat{h}) \leq \epsilon\right) \geq 1 - \delta$$

$$\epsilon < 1/2 \qquad \delta < 1/2$$

- Efficient PAC learning: polynomial in $1/\epsilon$, $1/\delta$

Assumptions:

- $\mathcal{H}$ is finite
- $\mathcal{H}$ contains hypotheses with $E(h) \leq \epsilon$
- Train and test examples from $\sim \mathcal{D}$

PAC Learning

Probably Approximately Correct Learning

- Consistent hypothesis: classifies training set with no error
- Version space $\mathcal{V}$: set of h s.t. $\hat{E}_S(h) = 0$
- A consistent hypothesis minimizes empirical error
- A consistent learner outputs hypotheses in $\mathcal{V}$
- Version space is ϵ-exausted if

$$\forall h \in \mathcal{V} \qquad E(h) < \epsilon$$

- $\mathcal{V}$ is not ϵ-exausted if

$$\exists h \in \mathcal{V} \qquad E(h) \geq \epsilon$$

- (Learner cannot tell this since it only encounters the training set)

PAC Learning

Probably Approximately Correct Learning

- Probability that no $h \in \mathcal{V}$ has $E(h) > \epsilon$?
- Consider $h_1, h_2, \ldots, h_k$ with $E(h) > \epsilon$
  - Probability of h consistent with one example $< 1 - \epsilon$
  - Probability of h consistent with m examples $< (1 - \epsilon)^m$
- P at least one $E(h) > \epsilon$ consistent with m examples $\leq k(1 - \epsilon)^m$

$$P(A_1 \cup A_2 \cup \ldots A_k) \leq P(A_1) + P(A_2) + \ldots + P(A_k)$$

- We don't know k, but since $k \leq |\mathcal{H}|$

$$k(1 - \epsilon)^m \leq |\mathcal{H}|(1 - \epsilon)^m$$

PAC Learning

Probably Approximately Correct Learning

- Since $(1-\epsilon) \leq e^{-\epsilon}$ for $0 < \epsilon < 1$:

$$k(1-\epsilon)^m \leq |\mathcal{H}|(1-\epsilon)^m \leq |\mathcal{H}|e^{-\epsilon m}$$

$$P\left(\exists h \in \mathcal{V} : E(h) \geq \epsilon\right) \leq |\mathcal{H}|e^{-\epsilon m}$$

- Upper bound on probability of not discarding all h with $E(h) > \epsilon$
- Lower bound on the number of examples for a consistent learner to learn an hypothesis with error below ϵ with a probability of $1-\delta$

$$P\left(E(\hat{h}) \leq \epsilon\right) \geq 1-\delta \qquad P\left(E(h \in \mathcal{V}) > \epsilon\right) \leq \delta \qquad m \geq \frac{1}{\epsilon}\left(\ln \frac{|\mathcal{H}|}{\delta}\right)$$

PAC Learning

Probably Approximately Correct Learning

- Upper bound on the error w.r.t. m with probability of $1-\delta$

$$P\left(E(\hat{h}) \leq \epsilon\right) \geq 1-\delta \qquad m \geq \frac{1}{\epsilon}\left(\ln\frac{|\mathcal{H}|}{\delta}\right) \Leftrightarrow \epsilon \leq \frac{1}{m}\left(\ln\frac{|\mathcal{H}|}{\delta}\right)$$

- This assumes $\hat{E}_S(\hat{h}) = 0$. Extending for $\hat{E}_S \geq 0$
- Training error is the mean of Bernoulli variables:

$$\hat{E}(h_i) = \frac{1}{m}\sum_{i=1}^{m} 1\{h(x^{(i)} \neq c(x^{(i)})\} = \frac{1}{m}\sum_{i=1}^{m} Z_i$$

- We can use Hoeffding inequalities:

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

$$P\left(E(h) > \hat{E}_S(h) + \epsilon\right) \leq e^{-2m\epsilon^2}$$

PAC Learning

Probably Approximately Correct Learning

$$P\left(E(h) > \hat{E}_S(h) + \epsilon\right) \leq e^{-2m\epsilon^2}$$

- But this is for one hypothesis. For all $h \in \mathcal{H}$:

$$P\left(\exists h \in \mathcal{H} : E(h) > \hat{E}_S(h) + \epsilon\right) \leq |\mathcal{H}|e^{-2m\epsilon^2}$$

- Calling this δ and solving for m:

$$m \geq \frac{1}{2\epsilon^2}(\ln \frac{|\mathcal{H}|}{\delta})$$

- Lower bound on $|S|$ to ensure generalization error below ϵ with confidence $1 - \delta$
- Increases quadratically with $1/\epsilon$ and with log of $|\mathcal{H}|$

PAC Learning

Inductive bias

- We mentioned that all learning algorithms must assume something about the function to learn (inductive bias). What if they don't?
- Example: let $\mathcal{H}$ be the set of all subsets of $\mathcal{X}$, so no inductive bias as it can represent any function $h : \mathcal{X} \rightarrow \{0, 1\}$
- Thus, $|\mathcal{H}| = 2^{|\mathcal{X}|}$

$$m \geq \frac{1}{2\epsilon^2}(\ln \frac{|\mathcal{H}|}{\delta}) \Leftrightarrow m \geq \frac{1}{2\epsilon^2}(\ln \frac{2^{|\mathcal{X}|}}{\delta}) \Leftrightarrow m \geq \frac{1}{2\epsilon^2}|\mathcal{X}| \ln \frac{2}{\delta}$$

- This requires that m be larger than $|\mathcal{X}|$, making learning impossible.

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

PAC Learning

Bias-Variance tradeoff

- What is the bound on generalization error for ERM hypothesis?

$$E(\hat{h}) - \hat{E}(\hat{h}) \qquad \hat{h} = \arg\min_{h \in \mathcal{H}} \hat{E}(h)$$

- Let h^* be the best possible hypothesis from $\mathcal{H}$:

$$h^* = \arg\min_{h \in \mathcal{H}} E(h)$$

- $P(E(\hat{h}) \leq \hat{E}(\hat{h}) + \epsilon) \geq 1 - \delta$
- Also, $\hat{E}(\hat{h}) \leq \hat{E}(h^*)$ and $E(h^*) \leq E(\hat{h})$, so
  $P(E(h^*) \leq \hat{E}(h^*) + \epsilon) \geq 1 - \delta$
- $P(E(\hat{h}) \leq E(h^*) + 2\epsilon) \geq 1 - \delta$

PAC Learning

Bias-Variance tradeoff

- Replacing, with $P = 1 - \delta$:

$$E(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} E(h)\right) + 2\sqrt{\frac{1}{2m} \ln \frac{|\mathcal{H}|}{\delta}}$$

- High bias, large $\min_{h \in \mathcal{H}} E(h)$
  - If this term dominates, we have underfitting
- High variance, large $|\mathcal{H}|$ and $2\sqrt{\frac{1}{2m} \ln \frac{|\mathcal{H}|}{\delta}}$
  - If this term dominates, we have overfitting

PAC Learning

Probably Approximately Correct Learning

- This assumes $|\mathcal{H}|$ is finite:

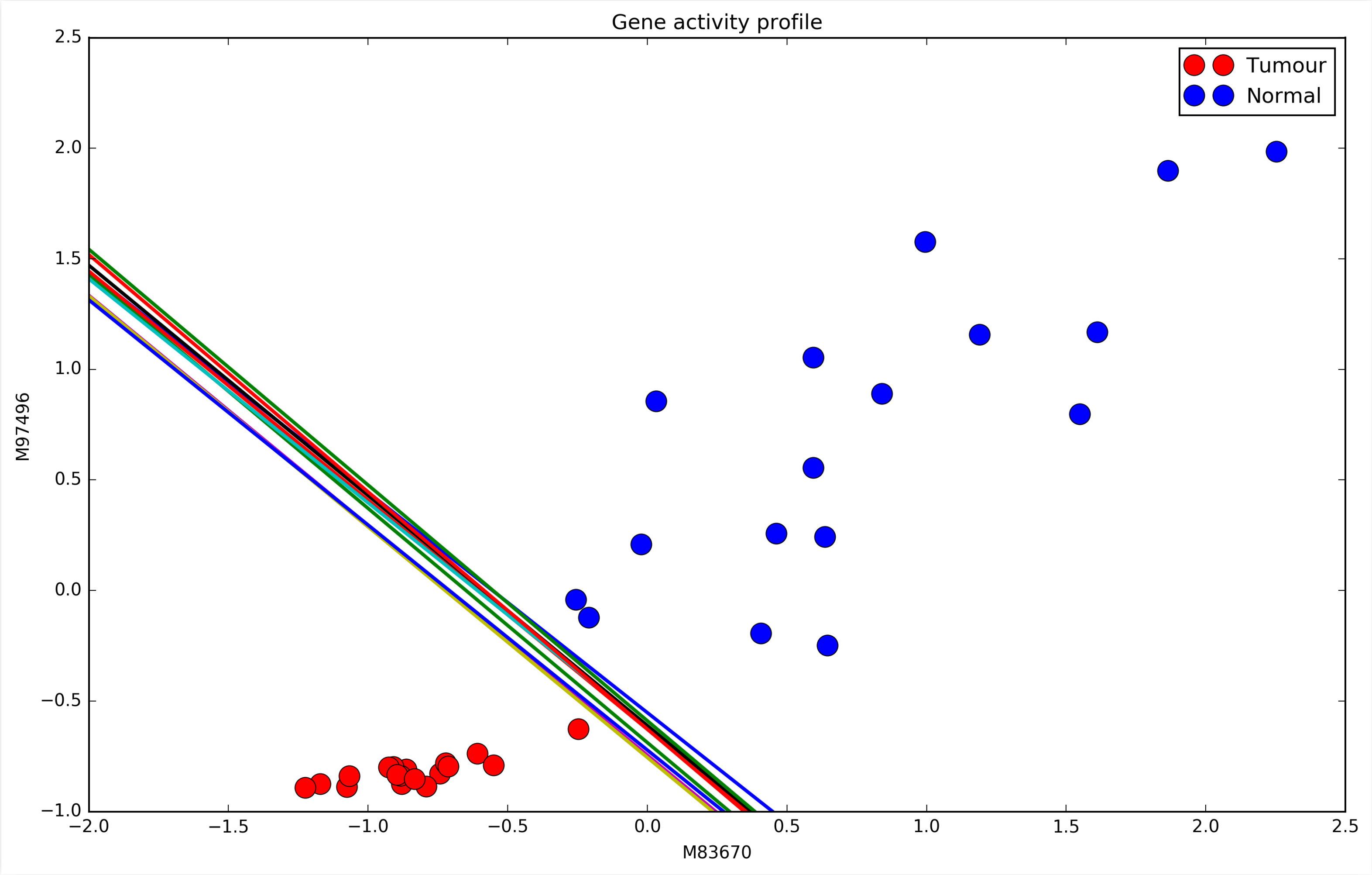
$$m \geq \frac{1}{2\epsilon^2}\left(\ln \frac{|\mathcal{H}|}{\delta}\right)$$

- True in some cases (e.g. limited-depth decision trees with categorical features) but false in general
- If $|\mathcal{H}|$ is infinite (e.g. discriminants with continuous parameters) then these limits are uninformative and we need a different approach

Shattering

Shattering

- Many hypotheses may be equivalent:

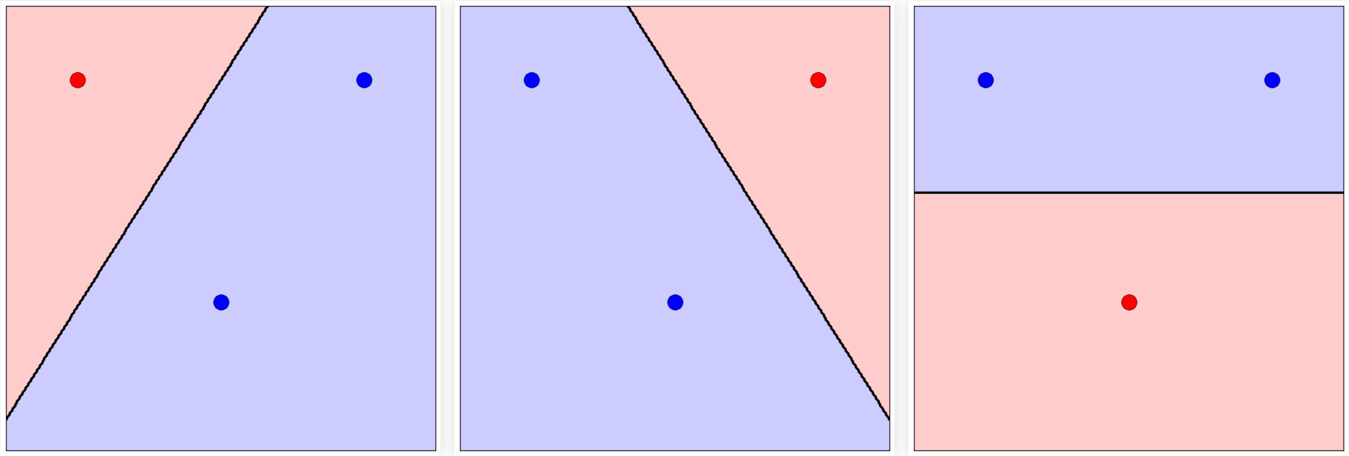


Shattering

- Instead of the total (infinite) number of hypotheses, we need some measure of how many hypotheses with different classification results the learner can generate

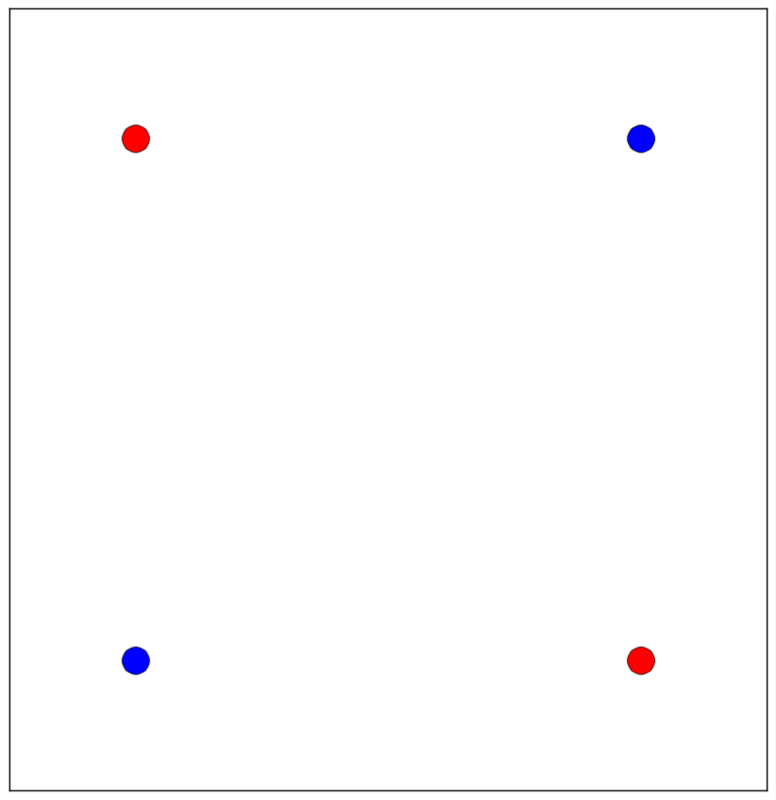
Shattering

- Hypothesis class $\mathcal{H}$ shatters set S if, for any labelling S, there is a $h \in \mathcal{H}$ consistent with S (classifies without errors)
- Example: linear classifier in 2D shatters 3 points



Shattering

- Example: linear classifier in 2D cannot shatter 4 points
  - There is no way to place 4 points such that all label combinations can be classified without error



V-C dimension

PAC Learning

V-C dimension

- The Vapnik-Chervonenkis dimension of $\mathcal{H}$, or $VC(\mathcal{H})$, is the size of the largest set S that $\mathcal{H}$ can shatter.

- There may be sets of size less than $VC(\mathcal{H})$ that cannot be shattered (e.g. two overlapping points, three points in a line, etc) but $VC(\mathcal{H})$ is the size of the largest that can be shattered

- From $VC(\mathcal{H})$, Vapnik et. al. demonstrated that, with $P \geq 1 - \delta$

$$E(\hat{h}) \leq \hat{E}(\hat{h}) + \mathcal{O}\left(\sqrt{\frac{VC(\mathcal{H})}{m} \ln \frac{m}{VC(\mathcal{H})} + \frac{1}{m} \ln \frac{1}{\delta}}\right)$$

- Roughly, size of training set must increase with $VC(\mathcal{H})$

PAC Learning

Linear discriminants

- We saw that we could increase the power of linear discriminants by increasing the number of dimensions
- We did this explicitely with logistic regression and saw how SVM do this implicitely with the kernel trick
- Linear discriminants of dimension D shatter D+1 points, so $VC(\mathcal{H}) = D + 1$
- Thus we can improve classification by increasing D
- But this also requires more data for training, otherwise overfitting

Summary

PAC Learning

Summary

- A solid statistical foundation provides useful intuitions
- although not used in practice; validation and test provide better estimates
- Inductive bias: necessary for learning so $|\mathcal{H}|$ not too large
- Bias-Variance tradeoff: best hypothesis vs $|\mathcal{H}|$
- Shattering and VC dimension for continuous models
- Results are not guaranteed, but only probably approximately correct

Further reading

- Mitchell, Chapter 7 up to section 7.4 (but outdated)
- Alpaydin, 2.1 - 2.3