Machine learning

Introduction to **Learning Theory**

Mohammad-Reza A. Dehaqani

dehaqani@ut.ac.ir

Overfitting:

- Huge feature space with kernels, what about **overfitting**?
- Maximizing margin leads to **sparse set** of support vectors
- Some **interesting theory** says that SVMs search for **simple** hypothesis with **large** margin and it is often **robust** to **overfitting**
- We have explored many ways of **learning from data**:
- But…
  - **How good** is our classifier, really?
  - **How much data** do I need to make it "good enough"?

How likely is a bad hypothesis to get m data points right?

- Classification with m i.i.d **data points** and **finite** number
- A learner finds a **hypothesis h** that is **consistent** with training data
  - Gets zero error in training: $error_{train}(h) = 0$ **($error_D(h)$)**
- The probability that h has **more** than ε error in test data ($error_{true}(h) \geq ε$ **($error_X(h) \geq ε$)**
- **Even** if h makes **zero** errors in **training data**, may make errors in **test**
- If $error_{true}(h) \geq ε$; the probability that h gets **one data point right**

$$\leq 1- ε$$

- Probability that h gets m data points **right**

$$\leq (1- ε)^m$$

How likely is a learner to pick a bad classifier?

- Usually there are many (say k) **bad** classifiers in **model class**

  $h_1, h_2, \ldots, h_k$ s.t. $\text{error}_{true}(h_i) \geq \varepsilon \quad i = 1, \ldots, k$

- Probability that learner **picks** a bad classifier

  = Probability that some bad classifier **gets 0 training error**

- Prob(h_1 gets 0 training error OR h_2 gets 0 training error OR … OR h_k gets 0 training error)

  $\leq$ Prob(h_1 gets 0 training error) + Prob(h_2 gets 0 training error) + … + Prob(h_k gets 0 training error)

  $\leq k\,(1\text{-}\varepsilon)^m$

- Probability that learner **picks a bad classifier**

  $\leq k\,(1\text{-}\varepsilon)^m \quad \leq |H|\,(1\text{-}\varepsilon)^m \quad \leq \quad |H|\,e^{-\varepsilon m}$

  Size of model class

PAC (Probably Approximately Correct) bound

- **Theorem [Haussler'88]**: **Model class H** finite, dataset D with m i.i.d. samples, $0 < \varepsilon < 1$: for any **learned classifier** h that gets **0 training error**:

$$P(\text{error}_{true}(h) \geq \epsilon) \leq |H|e^{-m\epsilon} \leq \delta$$

- Equivalently, with probability $\geq 1\text{-}\delta$

$$\text{error}_{true}(h) \leq \epsilon$$

- Important: **PAC bound** holds for all h with 0 training error, but **doesn't** guarantee that algorithm **finds best** h

Using a PAC bound

- Typically, 2 use cases:
  - Pick ε and δ, give you m
  - Pick m and δ, give you ε

$$|H|e^{-m\epsilon} \leq \delta$$

- Given ε and δ, yields **sample complexity**

Number of training data:

$$m \geq \frac{\ln|H| + \ln\frac{1}{\delta}}{\epsilon}$$

- Given m and d, yields **error bound**

Error:

$$\epsilon \geq \frac{\ln|H| + \ln\frac{1}{\delta}}{m}$$

Limitations of Haussler's bound

$$P(\text{error}_{true}(h) \geq \epsilon) \leq |H|e^{-m\epsilon}$$

- **Consistent** classifier:
  - Only consider classifiers with 0 training error
    - $\text{error}_{\text{train}}(\text{h}) = 0$

- **Size** of hypothesis space

$$m \geq \frac{\ln|H| + \ln\frac{1}{\delta}}{\epsilon}$$

  - what if |H| **too big** or H is **continuous** (e.g. linear classifiers)?

What if our classifier does not have zero error on the training data?

- A learner with zero training errors may make mistakes in test set
- What about a learner with $\mathrm{error}_{\mathrm{train}}(h) \neq 0$ in training set? ($\mathrm{error}_{\mathrm{train}}$ ($\mathrm{error}_{D}(h)$) relates $\mathrm{error}_{\mathrm{true}}$($\mathrm{error}_{X}(h)$))
- The error of a classifier is a **Bernoulli** random variable
  - like estimating the **parameter of a coin**!

$$\mathrm{error}_X(h) = \mathrm{error}_{\mathrm{true}}(h) := P(h(X) \neq Y) \equiv P(H=1) =: \theta$$

$$\mathrm{error}_D(h) = \mathrm{error}_{\mathrm{train}}(h) := \frac{1}{m}\sum_i \mathbf{1}_{h(X_i)\neq Y_i} \equiv \frac{1}{m}\sum_i x_i$$

Hoeffding's bound

- Consider m i.i.d. **Bernoulli random** variable $\mathbf{x}_1,\ldots,\mathbf{x}_m$, ($\mathbf{x}_i \in \{0,1\}$; flip a coin with **parameter** θ) . For $0<\varepsilon<1$:

$$P\left(\left|\theta - \frac{1}{m}\sum_i x_i\right| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

- For **a** single classifier h_i

$$P\left(|\text{error}_{true}(h_i) - \text{error}_{train}(h_i)| \geq \epsilon\right) \leq 2e^{-2m\epsilon^2}$$

- For **any** learned classifier $h \in H$ (we are comparing $|H|$ classifiers)

Union bound

$$P\left(|\text{error}_{true}(h) - \text{error}_{train}(h)| \geq \epsilon\right) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

Probability of mistake $\leq 2|H|e^{-2m\epsilon^2}$

Recall tail bounds in probability
Hoeffding's inequality (1963)

- The bounds when there are the range of the variables, but not the variances.

$$\left.\begin{array}{r} X_1, \ldots, X_n \text{ independent} \\ X_i \in [a_i, b_i] \\ \varepsilon > 0 \end{array}\right\} \Rightarrow$$

$$\Rightarrow \begin{cases} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| > \varepsilon\right) \leq 2\exp\left(\frac{-2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right) \\ \text{two-sided} \\ \\ \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) > \varepsilon\right) \leq \exp\left(\frac{-2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2}\right) \\ \text{one-sided} \end{cases}$$

PAC bound and Bias-Variance tradeoff

test set mistake

$$P\left(|\mathrm{error}_{true}(h) - \mathrm{error}_{train}(h)| \geq \epsilon\right) \leq 2|H|e^{-2m\epsilon^2} \leq \delta$$

- After moving some terms around, at least with **probability 1-δ:**

test set mistake — Bias (training mistake) — variance

$$\mathrm{error}_{true}(h) \leq \mathrm{error}_{train}(h) + \sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{2m}}$$

- **Fixed m**

Model Class		
Complex (|H| is Big)	Small	Large
Simple (|H| is small)	Large	Small

The size of the model class

$$2|H|e^{-2m\epsilon^2} \leq \delta$$

- Sample complexity

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

- How large is the model class?

$|H|$ is large $\Longrightarrow$ need many training examples

What about continuous hypothesis spaces

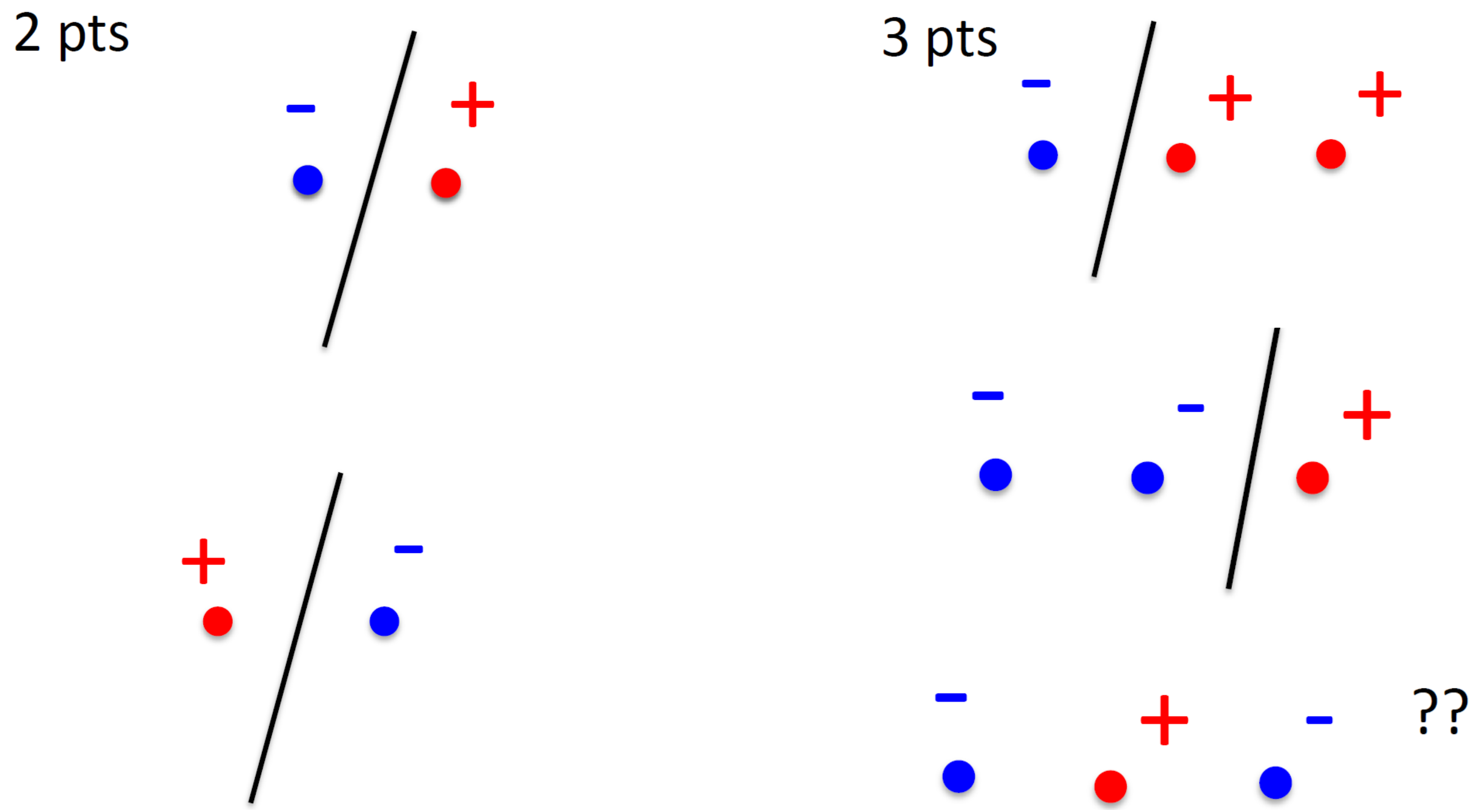
$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{2}{\delta}}{2m}}$$

- Continuous model class (e.g. linear classifiers):

  $|\text{H}| = \infty$

- Complexity of **model class** can depends on **maximum number of points** that can be **classified** exactly (and not **necessarily its size**)

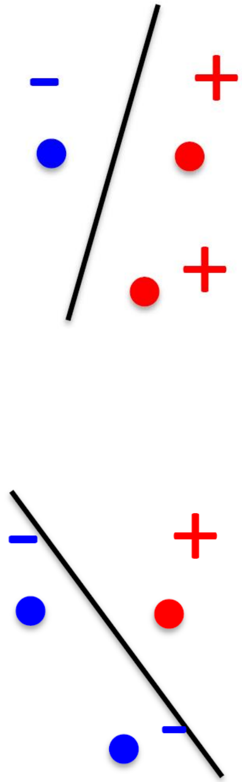
How many points can a linear boundary classify exactly? (1-D)



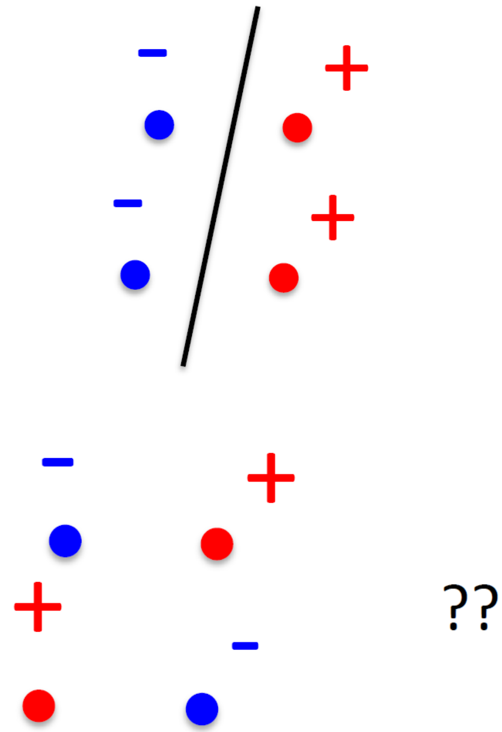
- There exists **placement** s.t. all labelings can be classified
  - Complexity of **model class is 2**

How many points can a linear boundary classify exactly? (2-D)

3 pts

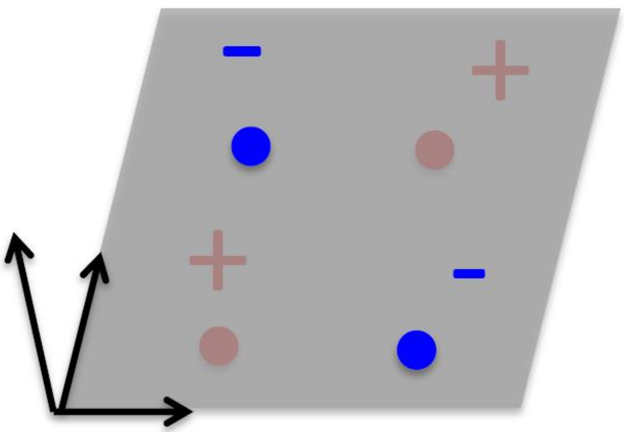


4 pts



- There exists **placement** s.t. all labelings can be classified
  - Complexity of **model class is 3**

How many points can a linear boundary classify exactly? (d-D)



- **Number** of training points that can be classified exactly is **VC dimension**

- How many parameters in linear Classifier in d-Dimensions?

$$w_0 + \sum_{i=1}^{d} w_i x_i$$

- d+1 parameters need d+1 **constrains**
- d+1 pts

PAC bound using VC dimension

- Measures relevant size of hypothesis space **using VC dimension**
- Bound for **infinite** dimension hypothesis spaces:
- In statistical learning, the Vapnik-Chervonenkis (VC) dimension is a popular measure of the **capacity** of a classifier.
- The **VC dimension** can predict a probabilistic upper bound on the **generalization error** of a classifier.

$$\sqrt{\frac{\ln|H| + \ln\frac{2}{\delta}}{2m}}$$

w.p. ≥ 1-δ

$$\text{error}_{true}(h) \le \text{error}_{train}(h) + \sqrt{\frac{VC(H)\left(\ln\frac{2m}{VC(H)} + 1\right) + \ln\frac{4}{\delta}}{m}}$$

Bias (training mistake)

variance

New game: Picking right VC(H)

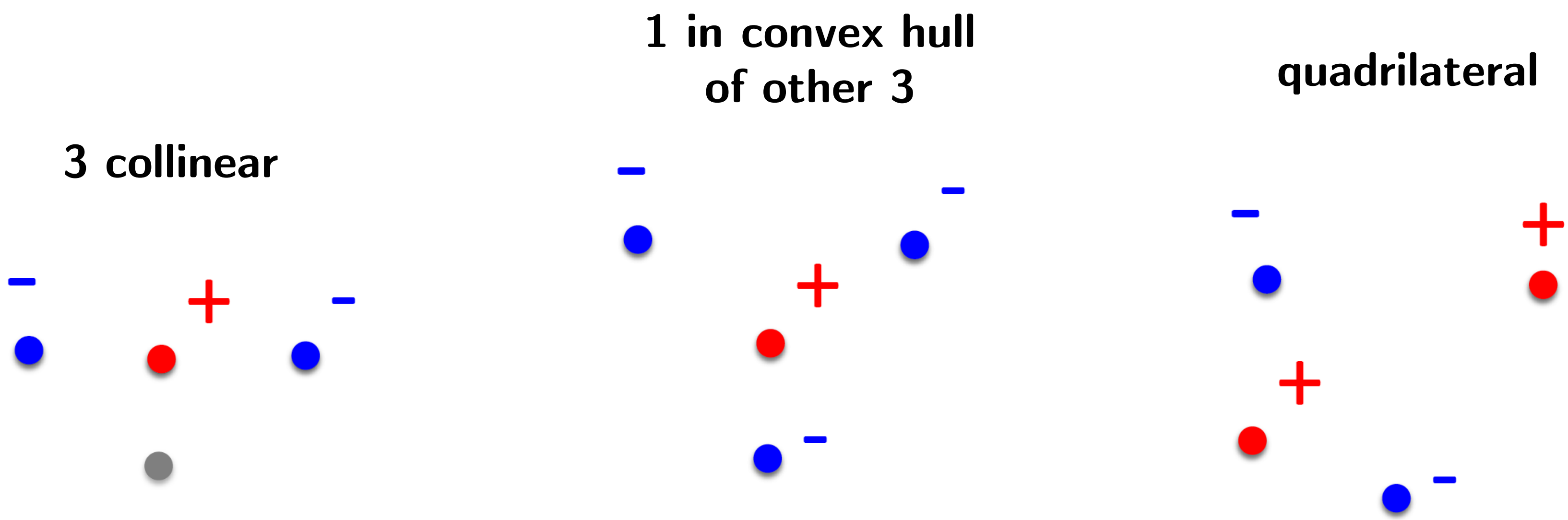
linear classifiers		
2D Small VC(H)	large	small
10,000 D Big VC(H)	small	large

VC (Vapnik-Chervonenkis) dimension

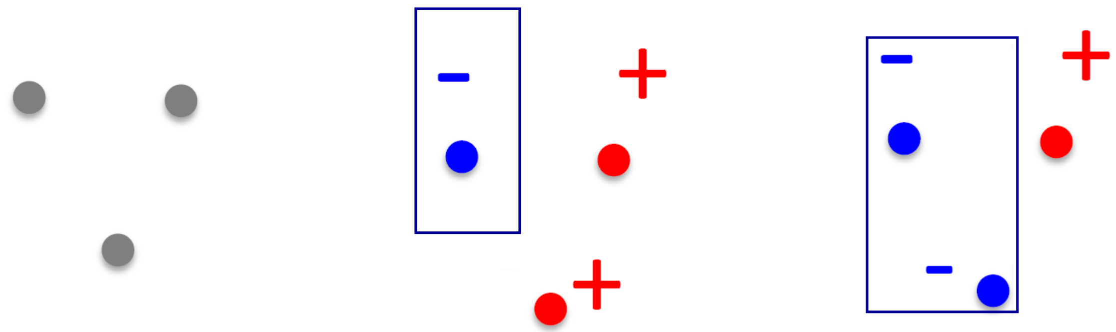
- Definition: VC dimension of a **hypothesis space H** is the **maximum number of points** such that **there exists a hypothesis** in H that is **consistent** with (can **correctly** classify) **any labeling** of the points.
    - You pick set of points
    - Adversary assigns labels
    - You find a hypothesis in H consistent with the labels
- If VC(H) = k, then for all k+1 points, there exists a labeling that **cannot be shattered** (can't find a hypothesis in H consistent with it)
- Definition: a set of instances S is **shattered** by hypothesis space H if and only if for **every dichotomy** (+/- labeling) of S there exists **some hypothesis** in H **consistent** with this dichotomy (labeling)
- VC(H), of hypothesis space H defined over instances space X is the size of the **largest** finite subset of X **shattered** by H.
- If the **arbitrarily large** finite sets of X can be shattered by H, then VC(H) $\equiv \infty$

VC dim. example – What can't we shatter

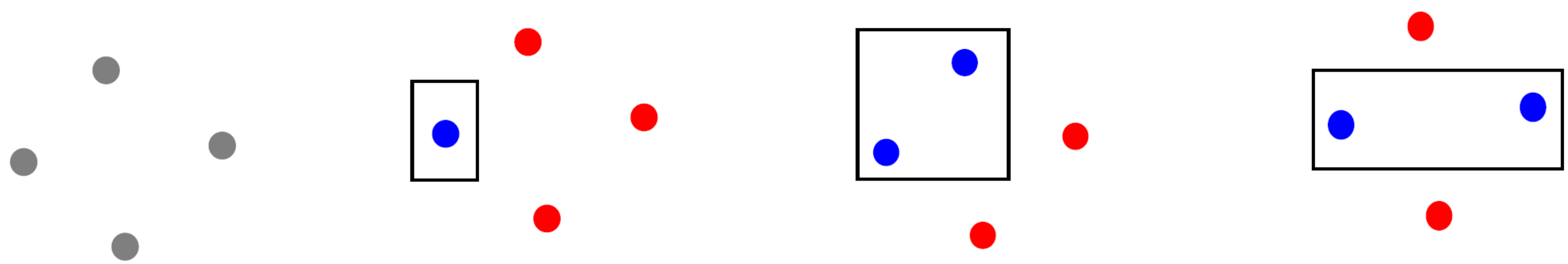
- What's the VC dim. of decision in 2D?
- If VC(H) = 3, then **for all placements of** 4 pts, there exists a labeling that can't be shattered
- **Linear classifiers**:
  - VC(H) = d+1, for *d* features plus constant term (3 if d=2)

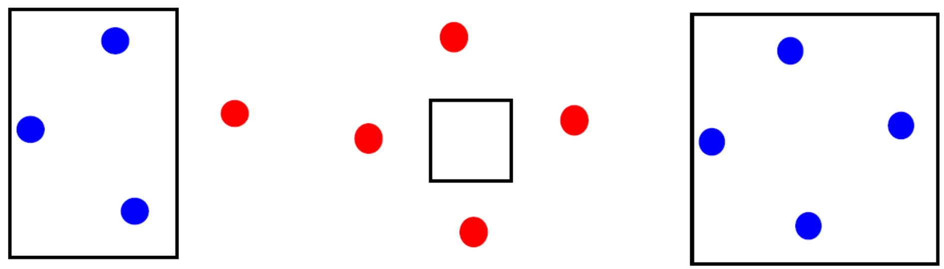


VC dim. of axis parallel rectangles in 2D



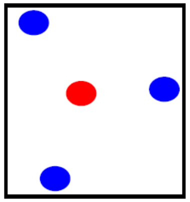
$VC(H) \geq 3$





$VC(H) \geq 4$

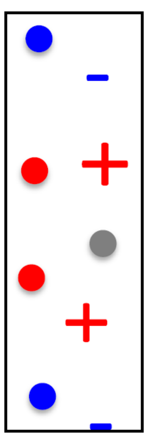
Some placement of 4 pts can't be **shattered**



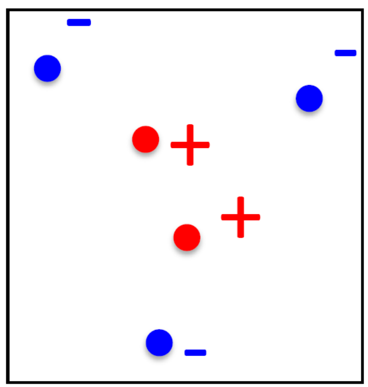
VC dim. of axis parallel rectangles in 2D is 4

- If VC(H) $= 4$, then for **all placements** of 5 pts, **there exists** a **labeling** that can't be shattered

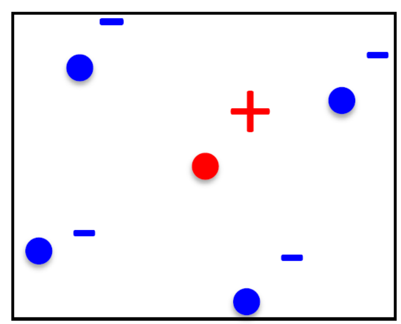
4 collinear



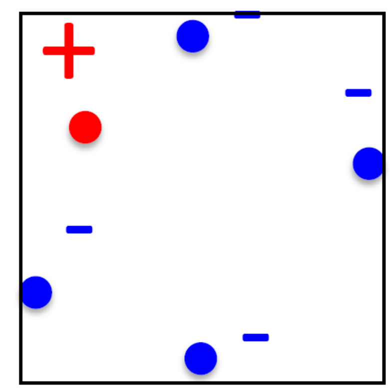
2 in convex hull of other 3



1 in convex hull of other 4



pentagon



Examples of VC dimension

- **Linear classifiers:**
  - VC(H) = d+1, for *d* features plus constant term
- **Axis parallel** rectangles:
  - VC(H) = 2d (4 if d=2)
- Nearest Neighbor:
  - VC(H) = ∞
- $VC(H) \leq \log_2|H|$ (So VC bound is tighter)
  - Given |H| hypothesis can hope to shatter max $m=\log_2|H|$ points
    - 2^m labelings $\Rightarrow$ $|H| \geq 2^m$

PAC bound for SVMs

- SVM uses **a linear classifier**
  - For d features, **VC(H) = d+1**:

$$\text{error}_{\mathcal{X}}(h) \leq \text{error}_{D}(h) + \sqrt{\frac{(d+1)\left(\ln \frac{2m}{d+1} + 1\right) + \ln \frac{4}{\delta}}{m}}$$

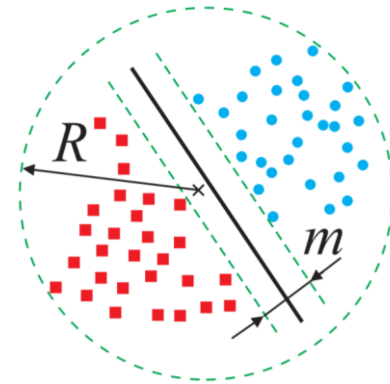
- **Polynomials** kernel:
  - Number of features grows really fast ⇒ Bad bound
  - Number of terms = $\binom{p+n-1}{p} = \frac{(p+n-1)!}{p!(n-1)!}$

n –input featuresp

p -degree of polynomial

- **Doesn't take margin into account**
- **Gaussian kernels** can classify any set of points exactly (VC(H) = ∞);
- Suggests Gaussian kernels and **deep nets (due to large number of parameters)** are really BAD!! But contradicts practice!

Margin-based VC dimension



- H: Class of linear classifiers: $\mathbf{w}.\Phi(\mathbf{x})$ (b=0)
  - Canonical form: $\min_j |\mathbf{w}.\Phi(\mathbf{x}_j)| = 1$
  - $VC(H) = R^2\mathbf{w.w}$
    - Doesn't depend on number of features!
    - $R^2 = \max_j \Phi(\mathbf{x}_j).\Phi(\mathbf{x}_j)$ –magnitude of data (R be the radius of the smallest inclosing ball (sphere) of the data)
    - R^2 is **bounded** even for Gaussian kernels $\rightarrow$bounded VC dimension
- **Large** margin $\Rightarrow$ **low** **w.w** $\Rightarrow$ low VC dimension $\Rightarrow$ low variance error
- SVMs minimize **w.w**
  - We require bound over **infinite number of possible** VC dimensions…

Structural risk minimization theorem

Bias (**empirical risk**)

$R^2\mathbf{w.w}$

$$\text{error}_{\mathcal{X}}(h) \leq \text{error}_{D}^{\gamma}(h) + C\sqrt{\frac{\frac{R^2}{\gamma^2}\ln m + \ln\frac{1}{\delta}}{m}}$$

Variance (**VC confidence**)

- γ is **margin**
- For a **family** of hyperplanes (with respective growing VC dimensions) with margin $\gamma>0$
- SVMs **maximize** margin γ+ **hinge loss**
  - Optimize **tradeoff** training error (bias) versus margin γ (variance)