

Lecture 9

Structural Risk Minimization

- Empirical Risk
- VC Dimension

Review of Lagrange multipliers

SVMs for the separable case

- Maximum margin hyper-plane
- The Lagrangian dual problem

Introduction

Learning a binary classification function from data

Given a dataset $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$ where each $y_i \in \{-1, 1\}$.
Learn a function $y = f(\mathbf{x}; \boldsymbol{\theta})$ that will correctly classify unseen examples.

How do we choose the type of f and $\boldsymbol{\theta}$?

By optimizing some measure of performance of the learned model.

What is a good measure of performance?

A good measure is the **expected risk**

$$R_f(\boldsymbol{\theta}) = \mathbb{E} [L(y, f(\mathbf{x}; \boldsymbol{\theta}))] = \text{Expected value of the Loss function}$$

Unfortunately, the risk cannot be measured directly since the underlying pdf is unknown. Instead, we typically use the risk over the training set, also known as the **empirical risk**

$$R_f^{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n L(y_i, f(\mathbf{x}_i; \boldsymbol{\theta})) = \text{Average value of loss function on the training set}$$

Introduction

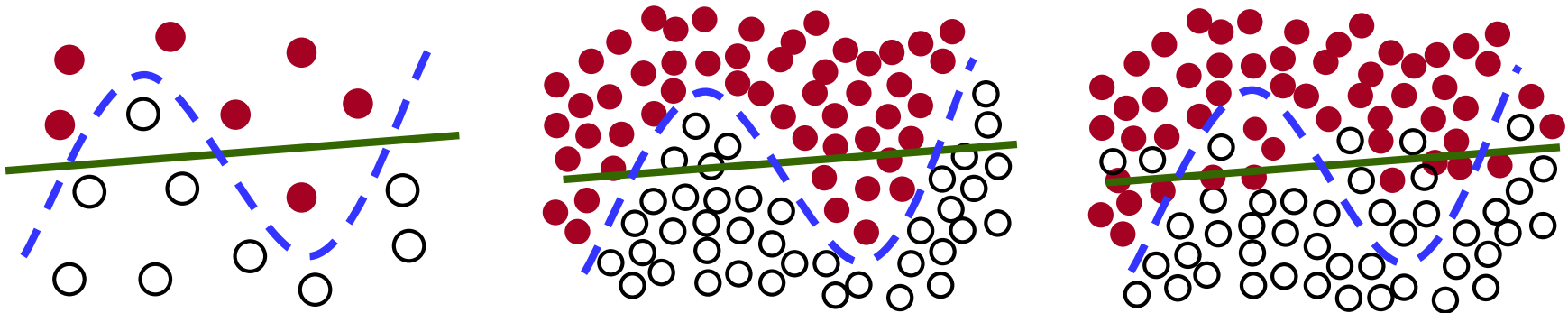
Empirical Risk Minimization

- A *formal* term for a *simple* concept: find the function $f(\mathbf{x})$ that minimizes the average risk on the training set.
- Minimizing the empirical risk is not a bad thing to do, provided that sufficient training data is available, since the law of large numbers ensures that the empirical risk will asymptotically converge to the expected risk for $n \rightarrow \infty$.
- However, for small samples, one cannot guarantee that ERM will also minimize the expected risk. This is the all too familiar issue of generalization.

Introduction

How do we avoid overfitting ?

By controlling model complexity. Intuitively, we should prefer the simplest model that explains the data (Occam's razor).



The VC dimension

The Vapnik-Chervonenkis dimension

This is a measure of the **complexity** / **capacity** of a class of functions \mathcal{F} . It measures the largest number of examples that can be explained by the family \mathcal{F} .

Trade-off between High Capacity and Good generalization

More capacity If the family \mathcal{F} has sufficient capacity to explain every possible data-set \implies there is a risk of overfitting.

Less capacity Functions $f \in \mathcal{F}$ having small capacity may not be able to explain our particular dataset, **however**, are much less likely to overfit.

How does VC-dimension characterize this trade-off?

Vapnik-Chervonenkis dimension

$$R_f(\boldsymbol{\theta}) = \mathbb{E} \left[\frac{1}{2} |y - f(\mathbf{x}; \boldsymbol{\theta})| \right] \quad R_f^{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{2} |y_i - f(\mathbf{x}_i; \boldsymbol{\theta})|$$

Given a class of functions \mathcal{F} , let h be its VC dimension.

h is a measure of \mathcal{F} 's capacity (h does not depend on the choice of training set)

Vapnik showed that with probability $1 - \eta$

$$R_f(\boldsymbol{\theta}) \leq R_f^{\text{emp}}(\boldsymbol{\theta}) + \sqrt{\frac{h(\log(\frac{2n}{h}) + 1) - \log(\frac{\eta}{4})}{n}}$$

This gives us a way to estimate the error on future data based only on the training error and the VC-dimension of \mathcal{F} .

Vapnik-Chervonenkis dimension

Given \mathcal{F} how do we define and compute h , its VC dimension?

Will now introduce the concept of shattering....

Shattering

A function $f(\mathbf{x}; \boldsymbol{\theta})$ can **shatter** a set of points $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_r$ if and only if

For every possible training set of the form $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_r, y_r)$, there exists some value of $\boldsymbol{\theta}$ such that $f(\mathbf{x}_i; \boldsymbol{\theta}) = y_i$ for $i = 1, \dots, r$.

Remember, there are 2^r such training sets to consider, each with a different combination of +1's and -1's for the y 's.

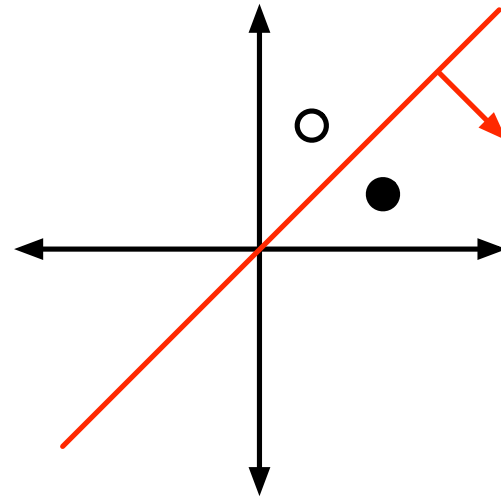
Shattering

A function f can **shatter** a set of points $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_r$ if and only if

For every possible training set of the form $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_r, y_r)$, there exists some value of θ such that $f(\mathbf{x}_i; \theta) = y_i$ for $i = 1, \dots, r$.

Question: Can the following f shatter the following points?

$$f(\mathbf{x}; \mathbf{w}) = \text{sgn}(\mathbf{w}^T \mathbf{x})$$

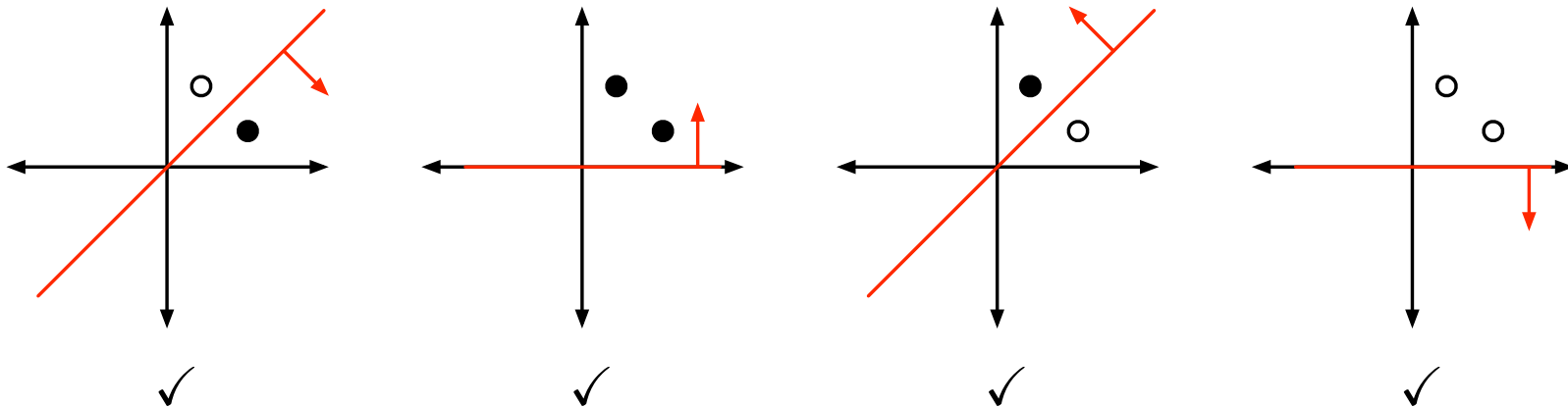


Shattering

A function f can **shatter** a set of points $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_r$ if and only if

For every possible training set of the form $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_r, y_r)$, there exists some value of θ such that $f(\mathbf{x}_i; \theta) = y_i$ for $i = 1, \dots, r$.

Answer: No problem. There are four training sets to consider



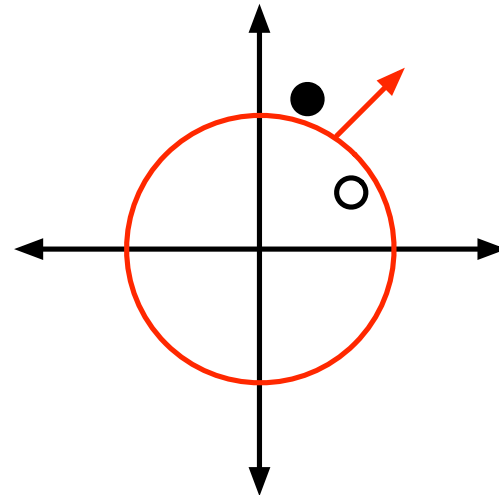
Shattering

A function f can **shatter** a set of points $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_r$ if and only if

For every possible training set of the form $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_r, y_r)$, there exists some value of θ such that $f(\mathbf{x}_i; \theta) = y_i$ for $i = 1, \dots, r$.

Question: Can the following f shatter the following points?

$$f(\mathbf{x}; b) = \text{sgn}(\mathbf{x}^T \mathbf{x} - b)$$

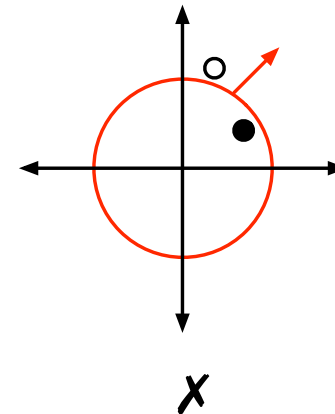
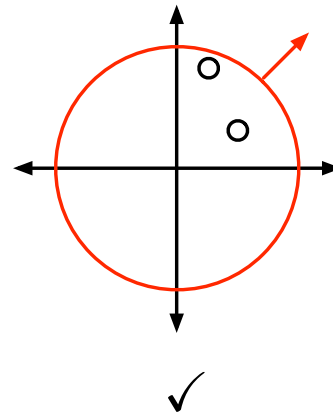
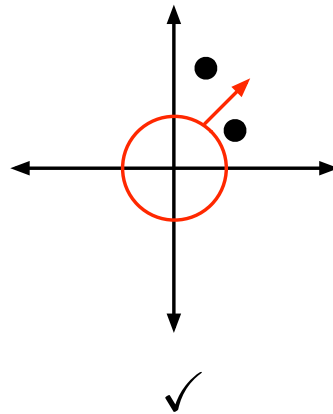
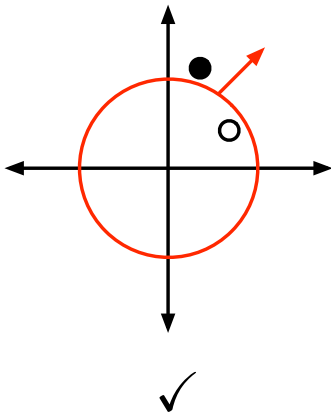


Shattering

A function f can **shatter** a set of points $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_r$ if and only if

For every possible training set of the form $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_r, y_r)$, there exists some value of θ such that $f(\mathbf{x}_i; \theta) = y_i$ for $i = 1, \dots, r$.

Answer: Not possible.



Definition of VC dimension

Given the class of functions \mathcal{F} , it has VC-dimension h if

there exists at least one set of h points that can be shattered by $f \in \mathcal{F}$ (note but, in general, it will not be true that every set of h points can be shattered).

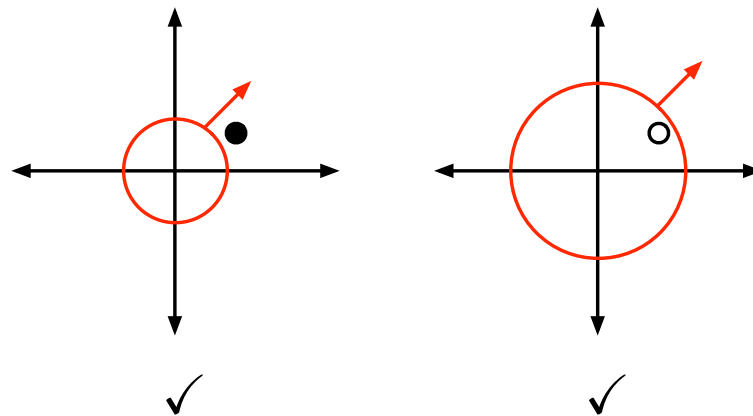
Question: What's the VC dimension of $f(\mathbf{x}; b) = \text{sgn}(\mathbf{x}^T \mathbf{x} - b)$?

Definition of VC dimension

Given the class of functions \mathcal{F} , it has VC-dimension h if

there exists at least one set of h points that can be shattered by $f \in \mathcal{F}$

Answer: 1. We can't even shatter two points! It's clear one point can be shattered.



Definition of VC dimension

Given the class of functions \mathcal{F} , it has VC-dimension h if

there exists at least one set of h points that can be shattered
by $f \in \mathcal{F}$

Example: For 2 dimensional inputs, what's the VC dimension of

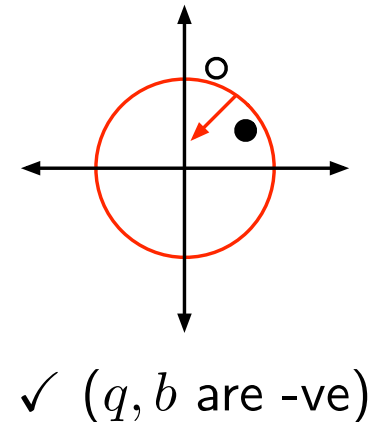
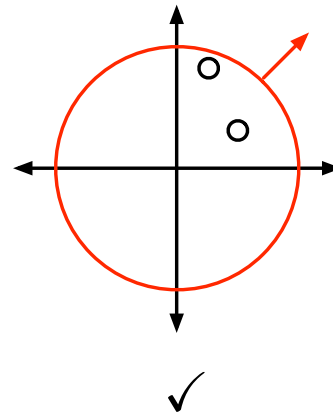
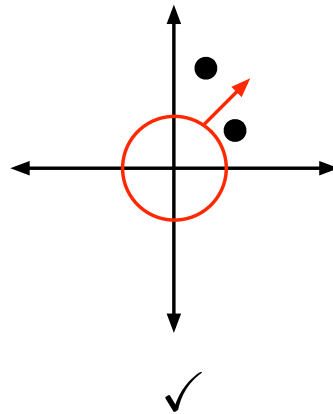
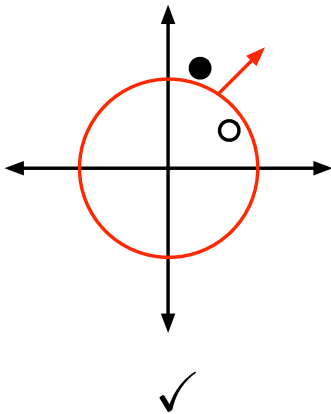
$$f(\mathbf{x}; q, b) = \text{sgn}(q \mathbf{x}^T \mathbf{x} - b)$$

Definition of VC dimension

Given the class of functions \mathcal{F} , it has VC-dimension h if

there exists at least one set of h points that can be shattered by $f \in \mathcal{F}$

Answer: 2



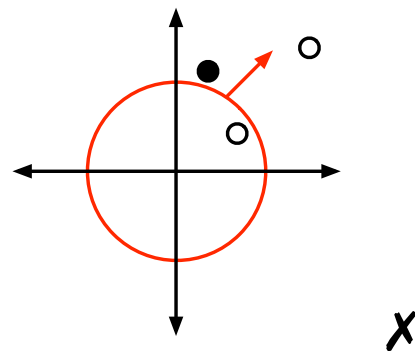
Definition of VC dimension

Given the class of functions \mathcal{F} , it has VC-dimension h if

there exists at least one set of h points that can be shattered by $f \in \mathcal{F}$

Example: What's the VC dimension of $f(\mathbf{x}; q, b) = \text{sgn}(q \mathbf{x}^T \mathbf{x} - b)$

Answer: 2 (clearly can't do 3)



VC dimension of separating line

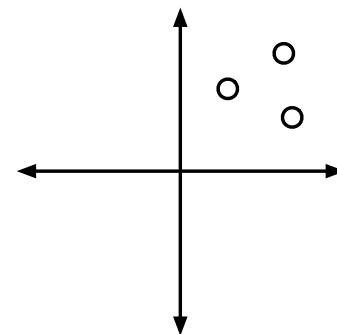
Given the class of functions \mathcal{F} , it has VC-dimension h if

there exists at least one set of h points that can be shattered by $f \in \mathcal{F}$

Example: For 2 dimensional inputs, what's the VC dimension of

$$f(\mathbf{x}; \mathbf{w}, b) = \text{sgn}(\mathbf{w}^T \mathbf{x} + b)$$

Can f shatter these 3 points?



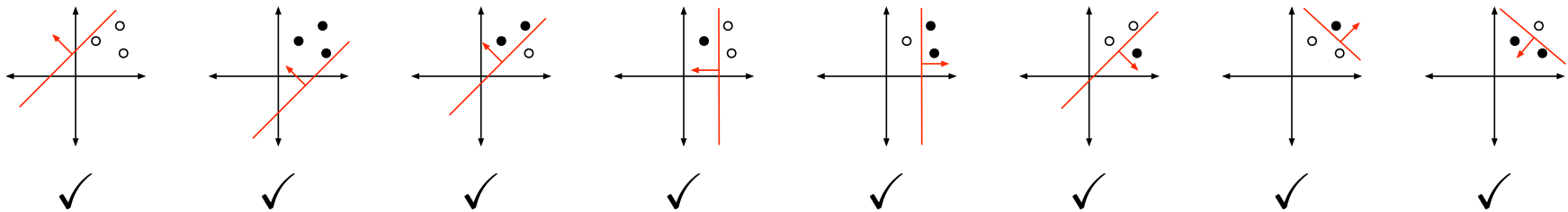
VC dimension of separating line

Given the class of functions \mathcal{F} , it has VC-dimension h if

there exists at least one set of h points that can be shattered by $f \in \mathcal{F}$

Example: What's the VC dimension of $f(\mathbf{x}; \mathbf{w}, b) = \text{sgn}(\mathbf{w}^T \mathbf{x} + b)$?

Answer: Yes, can shatter 3 points.



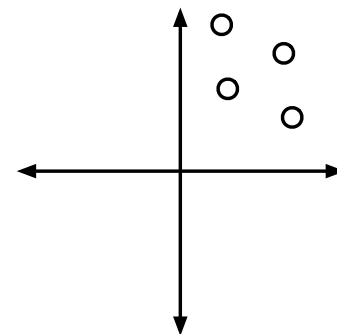
VC dimension of separating line

Given the class of functions \mathcal{F} , it has VC-dimension h if

there exists at least one set of h points that can be shattered by $f \in \mathcal{F}$

Example: For 2-dimensional inputs, what's the VC dimension of $f(\mathbf{x}; \mathbf{w}, b) = \text{sgn}(\mathbf{w}^T \mathbf{x} + b)$?

Can we find four points that f can shatter?



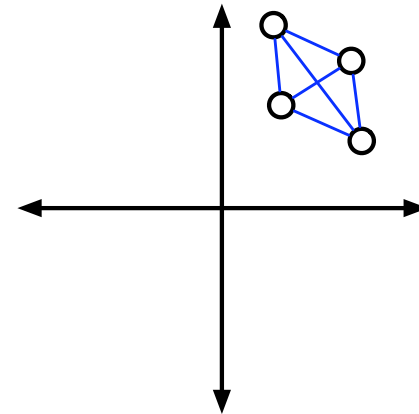
VC dimension of separating line

Given the class of functions \mathcal{F} , it has VC-dimension h if

there exists at least one set of h points that can be shattered by $f \in \mathcal{F}$

Example: What's the VC dimension of $f(\mathbf{x}; \mathbf{w}, b) = \text{sgn}(\mathbf{w}^T \mathbf{x} + b)$?

Can always draw 6 lines between pairs of four points.



VC dimension of separating line

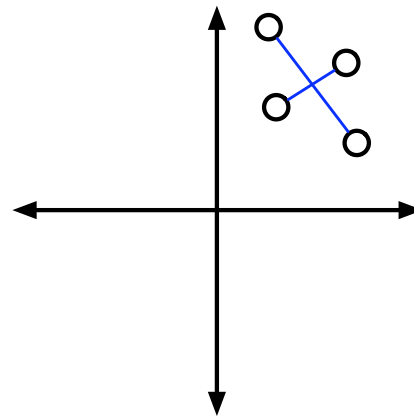
Given the class of functions \mathcal{F} , it has VC-dimension h if

there exists at least one set of h points that can be shattered by $f \in \mathcal{F}$

Example: What's the VC dimension of $f(\mathbf{x}; \mathbf{w}, b) = \text{sgn}(\mathbf{w}^T \mathbf{x} + b)$?

Can always draw 6 lines between pairs of four points.

Two of those lines will cross.



VC dimension of separating line

Given the class of functions \mathcal{F} , it has VC-dimension h if

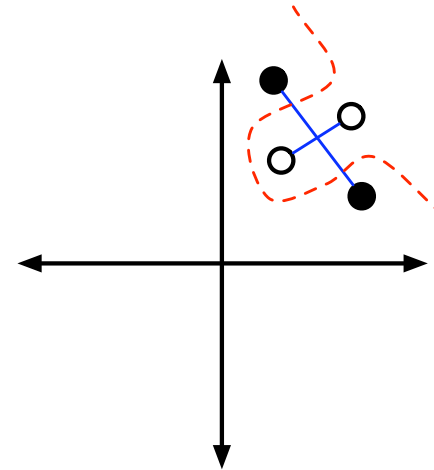
there exists at least one set of h points that can be shattered by $f \in \mathcal{F}$

Example: What's the VC dimension of $f(\mathbf{x}; \mathbf{w}, b) = \text{sgn}(\mathbf{w}^T \mathbf{x} + b)$?

Can always draw 6 lines between pairs of four points.

Two of those lines will cross.

If we put points linked by the crossing lines in the same class they can't be linearly separated.



A line can shatter 3 points but not 4 \implies VC-dim of a separating line is 3.

VC dimension of linear classifier in d -dimensions

If the input space is d -dimensional and if f is $\text{sgn}(\mathbf{w}^T \mathbf{x} - b)$, what is its VC-dimension?

Proof:

First will show d points can be shattered...

VC dimension of linear classifier in d -dimensions

If the input space is d -dimensional and if f is $\text{sgn}(\mathbf{w}^T \mathbf{x} - b)$, what is its VC-dimension?

Proof:

Define d input points thus:

$$\mathbf{x}_1 = (1, 0, 0, \dots, 0), \quad \mathbf{x}_2 = (0, 1, 0, \dots, 0), \quad \dots, \quad \mathbf{x}_d = (0, 0, 0, \dots, 1)$$

So $\mathbf{x}_{k,j} = 1$ if $k = j$ and 0 otherwise.

Let y_1, y_2, \dots, y_d be any one of the 2^d combinations of class labels.

How can we define \mathbf{w} and b to ensure

$$\text{sgn}(\mathbf{w}^T \mathbf{x}_k + b) = y_k \quad \text{for all } k ?$$

Remember

$$\operatorname{sgn}(\mathbf{w}^T \mathbf{x}_k + b) = \operatorname{sgn}\left(b + \sum_{i=1}^d x_{k,i} w_i\right) = \operatorname{sgn}(b + w_k)$$

Set $b = 0$ and $w_k = y_k$ for all k . Thus

$$f(\mathbf{x}_k; \mathbf{w}, b) = \operatorname{sgn}(\mathbf{w}^T \mathbf{x}_k + b) = \operatorname{sgn}(b + w_k) = \operatorname{sgn}(0 + y_k) = y_k$$

Thus the VC-dimension is $\geq d$.

Next:

Show that f can shatter $d + 1$ points. **How?**

As before set

$$\mathbf{x}_1 = (1, 0, 0, \dots, 0), \quad \mathbf{x}_2 = (0, 1, 0, \dots, 0), \quad \dots, \quad \mathbf{x}_d = (0, 0, 0, \dots, 1)$$

and additionally $\mathbf{x}_{d+1} = \mathbf{0} = (0, 0, 0, \dots, 0)$.

This time round set

$$\mathbf{w} = (y_1 - y_{d+1}, y_2 - y_{d+1}, \dots, y_d - y_{d+1}) \text{ and } b = y_{d+1}$$

Then for $k = 1, \dots, d$:

$$f(\mathbf{x}_k; \mathbf{w}, b) = \text{sgn}(\mathbf{w}^T \mathbf{x}_k + b) = \text{sgn}(b + w_k) = \text{sgn}(y_{d+1} + (y_k - y_{d+1})) = y_k$$

and for $k = d + 1$:

$$f(\mathbf{x}_{d+1}; \mathbf{w}, b) = \text{sgn}(\mathbf{w}^T \mathbf{x}_{d+1} + b) = \text{sgn}(b) = \text{sgn}(y_{d+1}) = y_{d+1}$$

Thus the VC-dimension is $\geq d + 1$.

Will prove in the next Exercise class that you cannot find a hyper-plane to shatter $d + 2$ points \implies VC dimension of the family of oriented separating hyperplanes in \mathcal{R}^d is $d + 1$.

What does VC-dimension measure?

Is it the number of parameters?

Related but not really the same

One may intuitively **expect** that models with a **large number of free parameters** would have **high VC** dimension, whereas models with **few parameters** would have **low VC** dimensions.

However, consider this example....

Example

Consider the one-parameter function

$$f_{\alpha}(x) = \text{sign}(\sin(\alpha x)), \quad x, \alpha \in \mathbb{R}.$$

Choose an arbitrary number h and set $x_i = 10^{-i}, i = 1, \dots, h$.

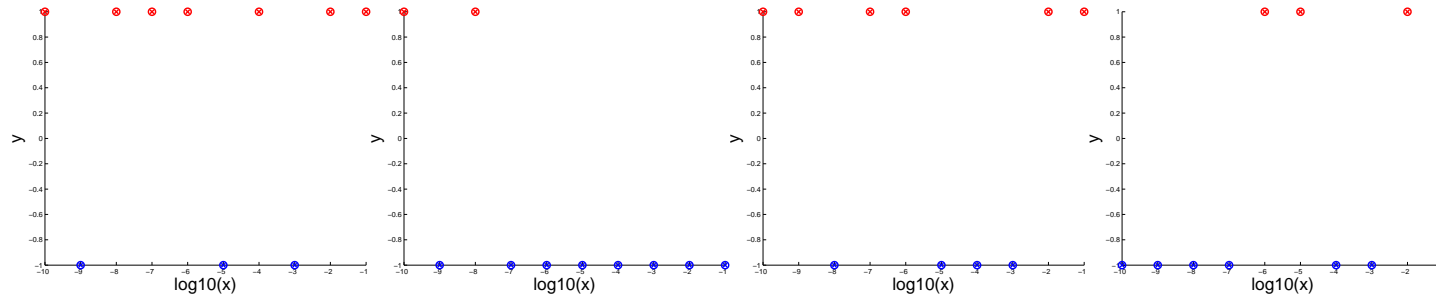
Choose the corresponding labels y_i arbitrarily with $y_i \in \{-1, +1\}$.

Let α be

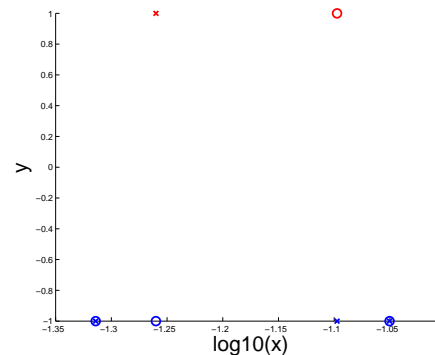
$$\alpha = \pi \left(1 + \sum_{i=1}^h \frac{(1 - y_i)10^i}{2} \right)$$

Despite having only one parameter, the function $f_{\alpha}(x)$ shatters an arbitrarily large number of points chosen according to the outlined procedure.

In this picture $h = 10$ and whatever the labelling predictions are correct. Circles the predictions and crosses are the ground truth.



But, can also find four points that cannot be shattered by this function!



So what do we make of this? The VC dimension is a more *sophisticated* measure of model complexity than dimensionality or number of free parameters [Pardo, 2000].

Structural Risk Minimization

Another formal term for an intuitive concept: the **optimal model** is found by **striking a balance** between the **empirical risk** and the **VC dimension**.

Remember:



















$$R(\boldsymbol{\theta}) \leq R^{\text{emp}}(\boldsymbol{\theta}) + \sqrt{\frac{h(\log(\frac{2n}{h}) + 1) - \log(\frac{\eta}{4})}{n}}$$

The SRM principle proceeds as follows

- Construct a nested structure for family of function classes $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subseteq \mathcal{F}_k$ with non-decreasing VC dimensions ($h_1 \leq h_2 \leq \dots \leq h_k$).
- For each class \mathcal{F}_i , compute solution f_i that minimizes the empirical risk.
- Choose the function class \mathcal{F}_i , and the corresponding solution f_i , that minimizes the risk bound.

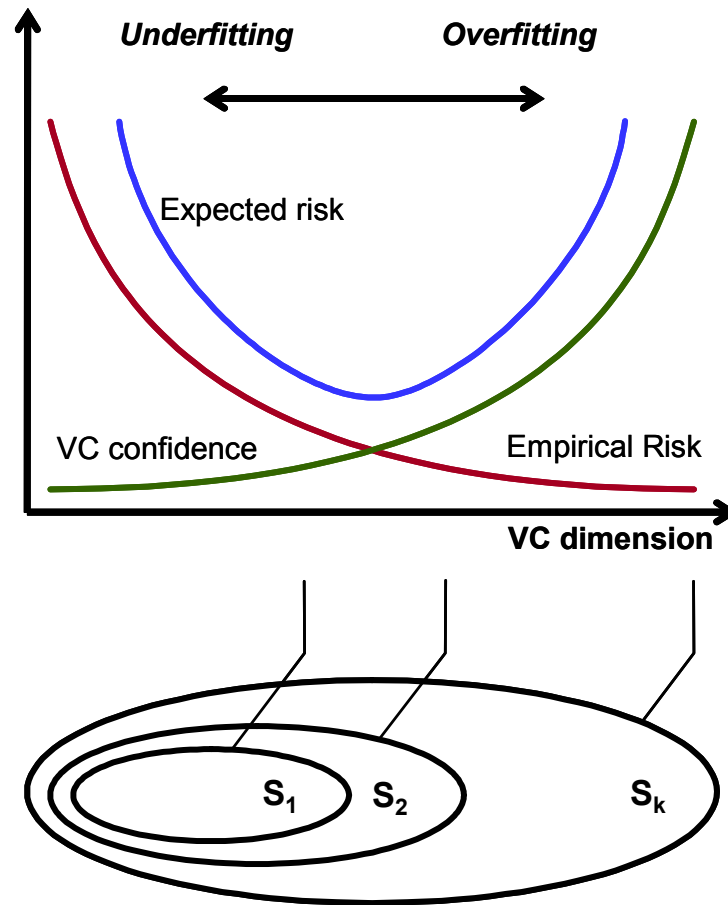
In other words

- Train a set of machines, one for each subset.
- For a given subset, train to minimize the empirical risk.
- Choose the machine whose sum of empirical risk and VC confidence is minimum.

| i | \mathcal{F}_i | $R^{\text{emp}}(\theta)$ | VC Confidence | Probable Upper bound | Choice |
|-----|-----------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|--------|
| 1 | \mathcal{F}_1 |  |  |  | |
| 2 | \mathcal{F}_2 |  |  |  | |
| 3 | \mathcal{F}_3 |  |  |  | |
| 4 | \mathcal{F}_4 |  |  |  | |
| 5 | \mathcal{F}_5 |  |  |  | |
| 6 | \mathcal{F}_6 |  |  |  | |

Note the second *VC-confidence* term is usually very, very conservative (at least hundreds of times larger than the empirical over-fitting effect).

Structural Risk Minimization



Using VC-dimensionality

People have worked hard to find the VC-dimension for













- Perceptrons
- Neural Nets
- Support Vector Machines
- And many many more

All with the goals of

1. Understanding which learning machines are more or less powerful under which circumstances.
2. Using *Structural Risk Minimization* to choose the best learning machine.

Alternatives to VC-dim based model selection

Could use potentially use **Cross-validation** instead:

| i | \mathcal{F}_i | $R^{\text{emp}}(\theta)$ | 10-Fold-CV-Error | Choice |
|-----|-----------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|--------|
| 1 | \mathcal{F}_1 |  |  | |
| 2 | \mathcal{F}_2 |  |  | |
| 3 | \mathcal{F}_3 |  |  | |
| 4 | \mathcal{F}_4 |  |  | |
| 5 | \mathcal{F}_5 |  |  | |
| 6 | \mathcal{F}_6 |  |  | |

Note the CV error might have more variance.

The VC dimension in practice

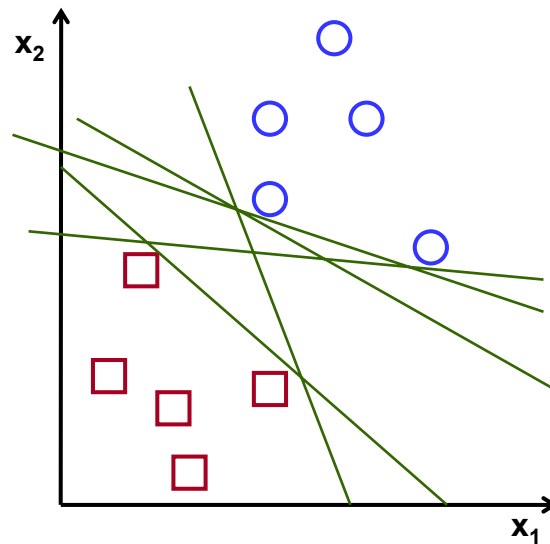
Unfortunately, computing an upper bound on the expected risk is not practical in various situations

- The VC dimension cannot be accurately estimated for non-linear models such as neural networks.
- Implementation of Structural Risk Minimization may lead to a non-linear optimization problem.
- The VC dimension may be infinite (e.g., $k = 1$ nearest neighbor), requiring infinite amount of data.
- The upper bound may sometimes be trivial (e.g., larger than one).

Fortunately, Statistical Learning Theory can be rigorously applied in the realm of linear models.

Optimal separating hyperplanes

Consider the problem of finding a separating hyperplane for a linearly separable dataset $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y \in \{-1, +1\}$.

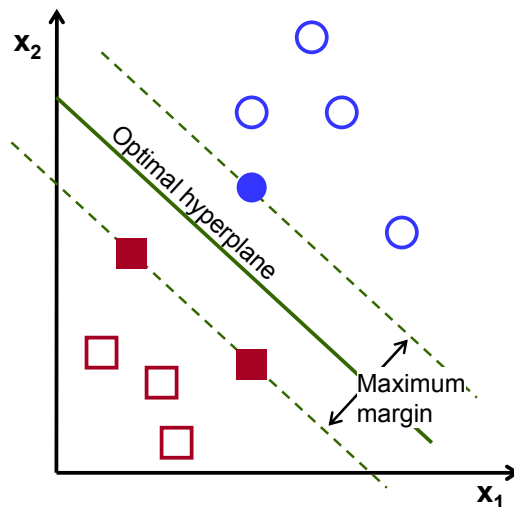


Which of the infinite hyperplanes should we choose?

Intuitively

Bad a hyperplane passing too close to the training examples will be sensitive to noise and probably less likely to generalize well

Better a hyperplane far away from all training examples will probably have better generalization capabilities.



Therefore, the optimal separating hyperplane will be the one with the largest **margin**, which is defined as the *minimum distance of an example to the decision surface*.

Optimal separating hyperplanes

How does this intuitive result relate to the VC dimension?

It can be shown [Vapnik, 1998] that the VC dimension of a separating hyperplane with a margin m is bounded as follows

$$h \leq \min \left(\frac{r^2}{m^2}, d \right) + 1$$

where d is the dimensionality of the input space, and r is the radius of the smallest sphere containing all the input vectors.

By **maximizing the margin** one is **minimizing the VC dimension**.

The separating hyperplane has zero empirical error and maximizing the margin \implies minimizing the upper bound on the expected risk.

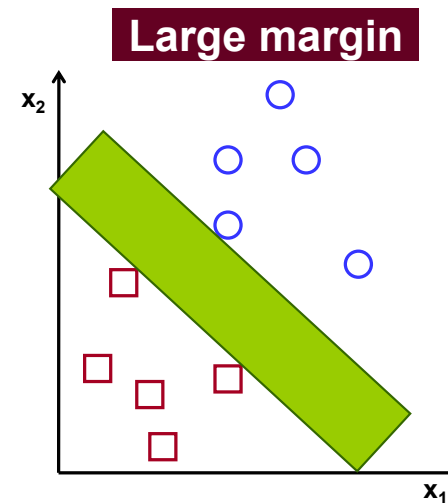
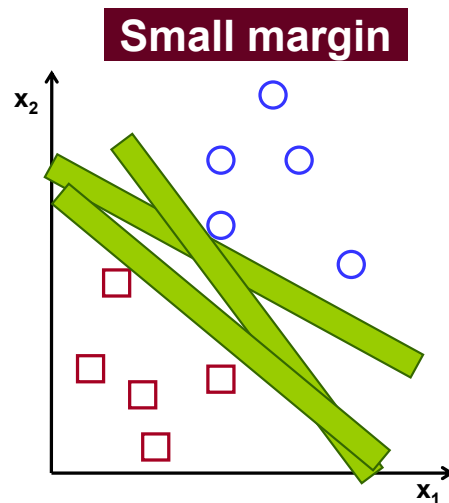
Therefore: The **separating hyperplane** with **maximum margin** will also **minimize the structural risk**.

Optimal separating hyperplanes

To further understand the relationship between margin and capacity, consider the two separating hyperplanes depicted below

A *skinny one* (small margin), will be able to adopt many orientations.

A *fat one* (large margin), will have limited flexibility.



A larger margin necessarily results in lower capacity

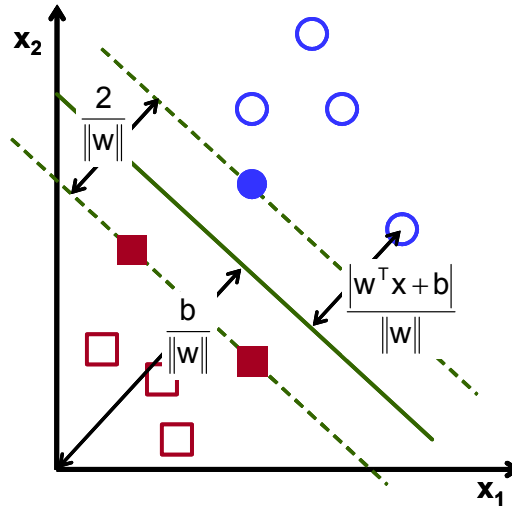
- We normally think of complexity as being a function of the number of parameters.

*Instead, Statistical Learning Theory tells us that if the **margin is sufficiently large**, the **complexity of the function will be low** even if the **dimensionality is very high!***

Optimal separating hyperplanes

Express the margin in terms of w and b of the separating hyperplane.

The distance between a point x and a plane (w, b) is $\frac{|w^T x + b|}{\|w\|}$



The optimal hyperplane has an infinite number of representations by simply re-scaling the weight vector and bias.

Choose the representation for which the discriminant function becomes **one** for the training examples closest to the boundary.

$$|\mathbf{w}^T \mathbf{x} + b| = 1, \quad \Leftrightarrow \text{the canonical hyperplane}$$

Therefore, the distance from the closest example to the boundary is

$$\frac{|\mathbf{w}^T \mathbf{x} + b|}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$$

The margin becomes

$$m = \frac{2}{\|\mathbf{w}\|}$$

Optimal separating hyperplanes

The problem of maximizing the margin is equivalent to

$$\text{minimize } J(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 \quad \forall i$$

Notice that $J(\mathbf{w})$ is a quadratic function, which means that there exists a single global minimum and no local minima.

To solve, use classical Lagrangian optimization techniques

The Karush-Kuhn-Tucker conditions are used to analyse the solution.

Lagrange multipliers

Our optimization problem is

$$\min_{\mathbf{w}} f(\mathbf{w}) \quad \text{subject to} \quad g_j(\mathbf{w}, b) \leq 0, \quad j = 1, \dots, n.$$

where

$$f(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{and} \quad g_j(\mathbf{w}, b) = 1 - y_j(\mathbf{w}^T \mathbf{x}_j + b)$$

As $f(\mathbf{x})$ is convex as is each $g_j(\mathbf{x})$ and we are **assuming the points are linearly separable**, we can invoke the theorem we stated earlier and solve the dual problem....

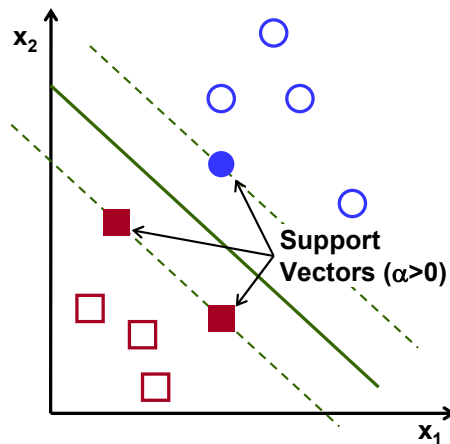
Properties of the solution

Karush-Kuhn-Tucker conditions: For an optimal feasible, $\mathbf{w}^*, b^*, \lambda^*$ solution the following conditions hold:

$$\text{KKT dual complementary condition} \rightarrow \lambda_i^* g_i(\mathbf{w}^*, b^*) = 0, \quad i = 1, \dots, n$$

$$g_i(\mathbf{w}^*, b^*) \leq 0, \quad i = 1, \dots, n$$

$$\lambda_i^* \geq 0, \quad i = 1, \dots, n$$



If $\lambda_i^* > 0$ then $g_i(\mathbf{w}^*, b^*) = 0 \implies$ the constraint g_i is **active**.

Thus the SVM in fact only depends only a small number of **support vectors**.

The dual problem

Our Lagrangian is

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{j=1}^n \lambda_j (1 - y_j (\mathbf{w}^T \mathbf{x}_j + b))$$

The Lagrange dual function of our optimization problem

$$\Theta(\boldsymbol{\lambda}) \equiv \min_{\mathbf{w}, b} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda}) = \mathcal{L}(\mathbf{w}^*, b^*, \boldsymbol{\lambda})$$

Finding \mathbf{w}^* and b^* requires computing the gradient

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda}) = \mathbf{w} - \sum_{j=1}^n \lambda_j y_j \mathbf{x}_j = \mathbf{0}$$

This implies that

$$\mathbf{w}^* = \sum_{j=1}^n \lambda_j y_j \mathbf{x}_j$$

Also

$$\frac{\partial \mathcal{L}(\mathbf{w}, b, \boldsymbol{\lambda})}{\partial b} = - \sum_{j=1}^n \lambda_j y_j = 0$$

Plugging \mathbf{w}^* and b^* into \mathcal{L} get

$$\mathcal{L}(\mathbf{w}^*, b^*, \boldsymbol{\lambda}) = \sum_{j=1}^n \lambda_j - b \sum_j \lambda_j y_j - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

but $\sum_j \lambda_j y_j = 0$, thus

$$\mathcal{L}(\mathbf{w}^*, b^*, \boldsymbol{\lambda}) = \sum_{j=1}^n \lambda_j - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

Putting everything together get the dual optimization problem

$$\max_{\boldsymbol{\lambda}} \left\{ \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right\}$$

subject to $\lambda_j \geq 0$ for $i = 1, \dots, n$ and $\sum_j \lambda_j y_j = 0$.

Properties of the solution

If \mathbf{w}^* found

$$b^* = -\frac{1}{2} \left(\max_{i:y_i=-1} \mathbf{w}^{*T} \mathbf{x}_i + \min_{i:y_i=1} \mathbf{w}^{*T} \mathbf{x}_i \right)$$

Prediction at a new point \mathbf{x} having found the optimal λ

$$\begin{aligned} \mathbf{w}^T \mathbf{x} + b &= \left(\sum_{i=1}^n \lambda_i y_i \mathbf{x}_i \right)^T \mathbf{x} + b \\ &= \sum_{i=1}^n \lambda_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \end{aligned}$$

Have to calculate a quantity that depends only on the inner product between \mathbf{x} and the points in the training set. The λ_i 's **will all be zero except for the support vectors**. Thus many of the terms in the sum zero. Only need to find the inner products between \mathbf{x} and the support vectors.

Pen & Paper assignment

- Details available on the course website.
- Your assignment is a small pen & paper exercise based on hand calculation of a separating line.
- Mail me about any errors you spot in the Exercise notes.
- I will notify the class about errors spotted and corrections via the course website and mailing list.