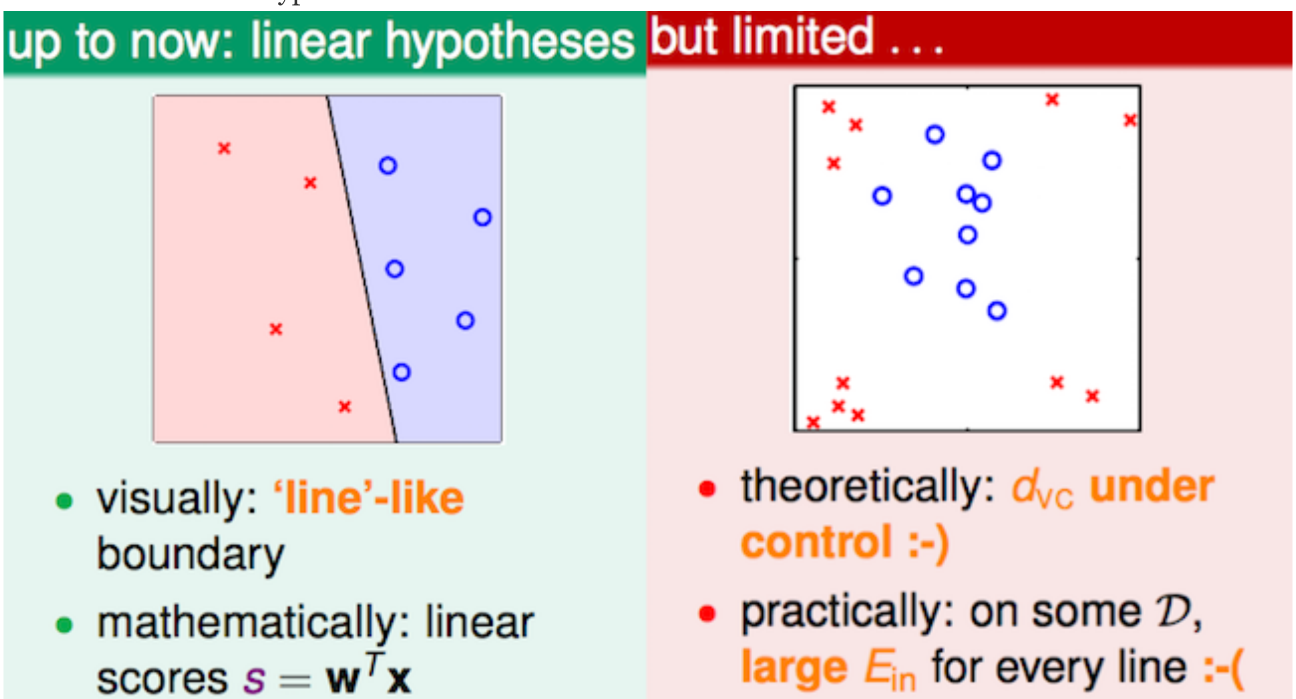
Week 12: Nonlinear Transformation

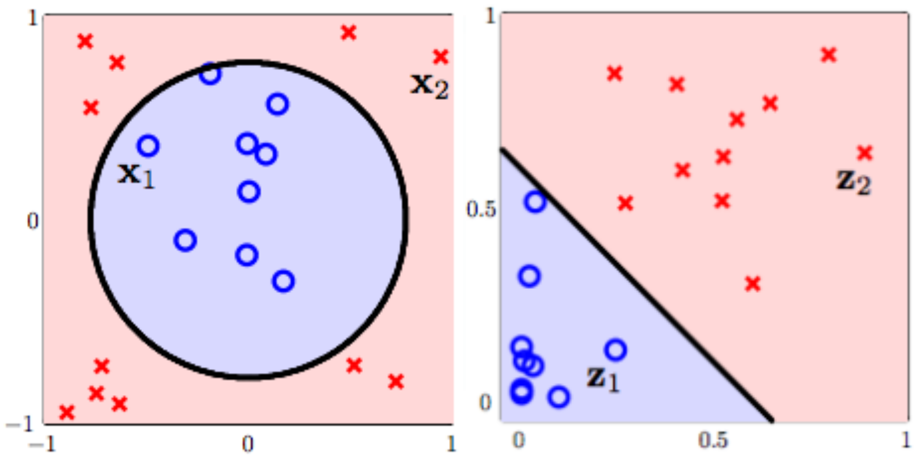
Table of Contents

Quadratic Hypothesis

1. Limitations of linear hypothesis



   - Linear hypothesis have small VC dimension and computational complexity
   - However, not every input set $\mathcal{D}$ can be learned well by linear hypothesis as is
2. Nonlinear feature transform
   - Map a **non linear-separable** set $\{(x_n, y_n)\}$ in χ space to **linear separable** $\{(z_n, y_n)\}$ in $\mathcal{Z}$ space
   - $x \in \chi \overset{\phi}{\mapsto} z \in \mathcal{Z}$: **(nonlinear) feature transform**

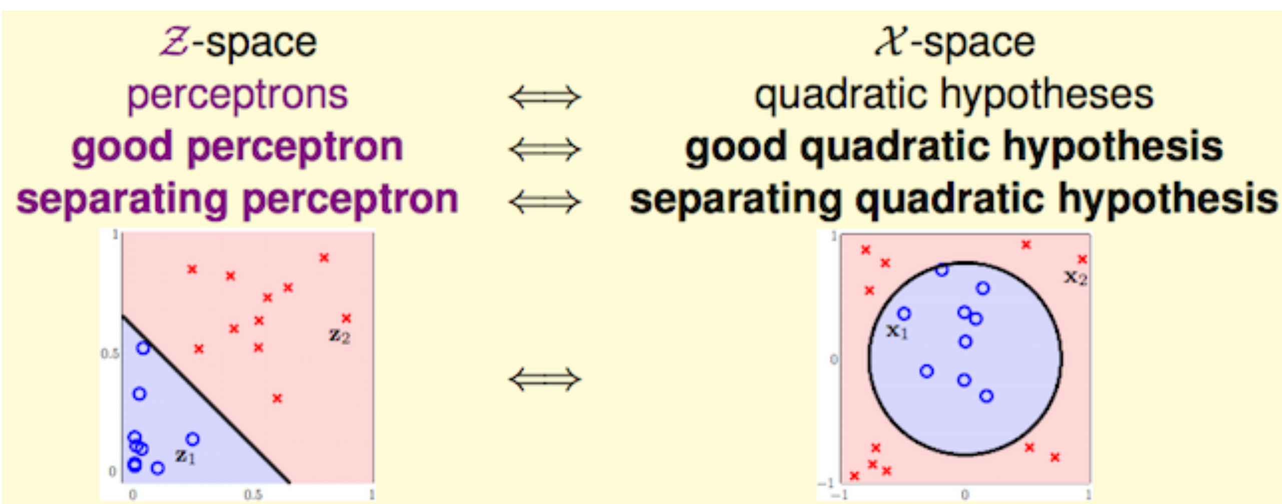


3. Linear hypotheses in $\mathcal{Z}$-space
    - Given a general (quadratic) $\mathcal{Z}$ space with $\phi(x) = (1, x_1, x_2, x_1^2, x1x2, x2^2)$, **perceptrons** in $\mathcal{Z}$-space ⇔ **quadratic hypotheses** in 2D χ-space

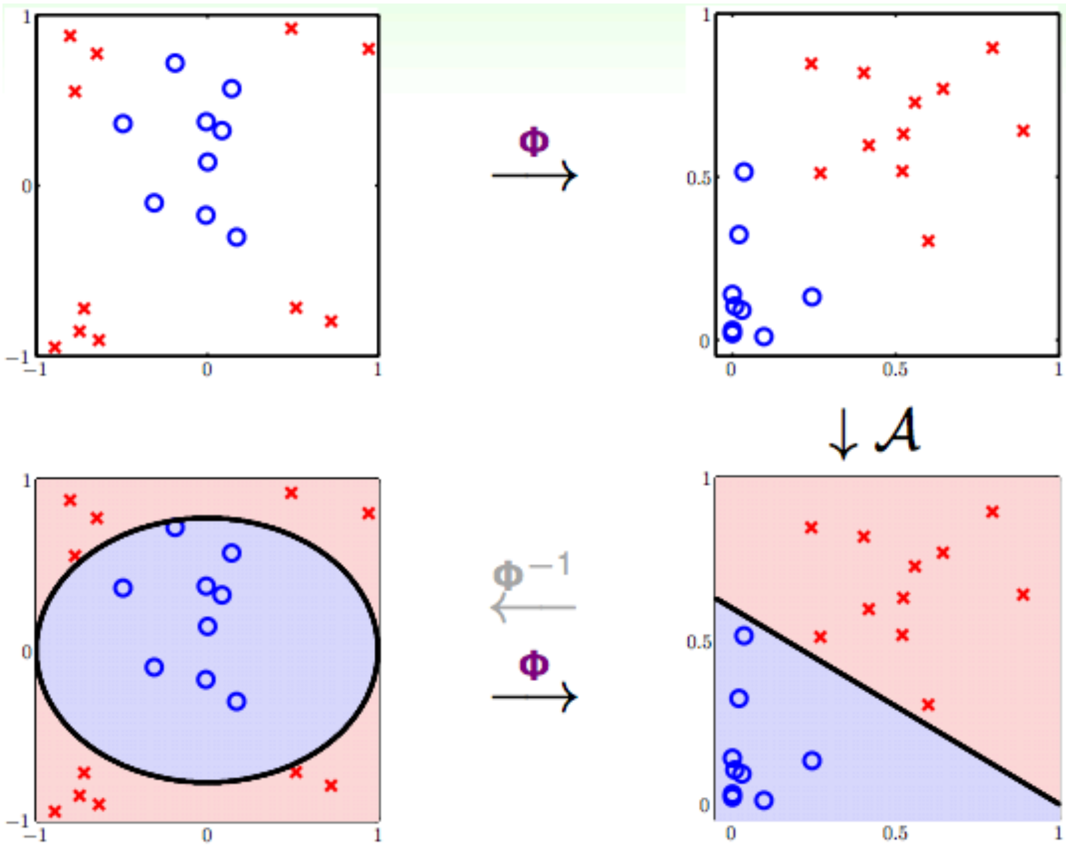
$$\mathcal{H}_\phi = \{h(x) : h(x) = \tilde{h}(\phi(x)) \text{ for some linear } \tilde{h} \text{ on } \mathcal{Z}\}$$

    - Such nonlinear transformation could implement **all possible quaratic curve boundaries** in $\mathcal{Z}$-space as **lines** in χ-space
        - Circle, ellipse, **rotated** ellipse, hyperbola, parabola
        - Including lines and constants **as degenrate cases**

Nonlinear Transform



1. Nonlinear transform: process of learning a **good perceptron** in $\mathcal{Z}$-space with data $\{(z_n = \phi(x_n), y_n)\}$
2. Nonlinear transform steps



- Transform original data $\{(x_n, y_n)\}$ to $\{(z_n = \phi(x_n), y_n)\}$ by ϕ
- Learn a good perceptron weight vector $\tilde{w}$ using $\{(z_n, y_n)\}$ and some linear classification algorithm $\mathcal{A}$
  - The learning happens in $\mathcal{Z}$-space
  - Can use any linear classification algorithm, i.e. PLA, linear regression, logistic regression
- Return $g(x) = sign(\tilde{w}^T \phi(x))$
  - On the surface, the reverse of step 1
  - But really, a process of mapping each training data **in χ-spacee**, to a corresponding point **in $\mathcal{Z}$-space**, and use the learned model to classify the mapped z_n

3. Both the **feature transform ϕ** and **linear model $\mathcal{A}$** are open to choices in nonlinear transform learning

Price of Nonlinear Transform

For $Q - thorder$ polynomial transform

$$\begin{aligned}\phi(x) = (1,& \\ & x_1, x_2, \dots, x_d, \\ & x_1^2, x1x2, \dots, x_d^2, \\ & \dots, \\ & x_1^Q, x_q^{Q-1}x2, \dots, x_d^Q)\end{aligned}$$

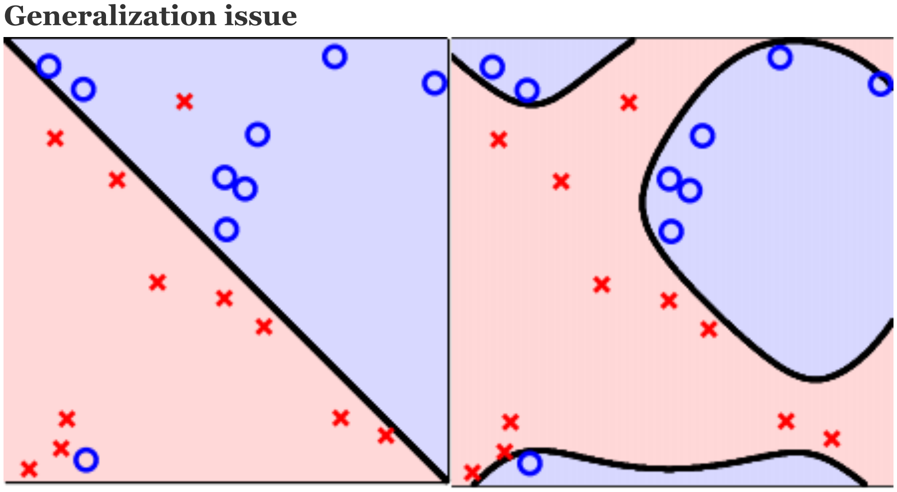
1. Computation/storage price

Efforts needed for computing/storing $z = \phi_Q(x)$ and $\tilde{w}$

$$
\begin{aligned}
&= \underbrace{1}_{\tilde{w}_0} + \underbrace{\tilde{d}}_{others} \quad \text{dimensions} \\
&= \text{Choose smaller or equal to Q items from d with repetition} \\
&= \binom{Q+d}{Q} \\
&= \binom{Q+d}{d} \\
&= O(Q^d)
\end{aligned}
$$

- Explanation for above equation
    - Total of x (and product of x) terms
    - Pick Q (up to order) or few than Q (special case, some terms are zero) from the set to form possible feature transforms
    - Total number of possible transforms is d choose Q **with replacement**

2. Model complexity price
    - For Q-th order polynomial transform with $\underbrace{1}_{\tilde{w}_0} + \underbrace{\tilde{d}}_{others}$ dimensions
        - Number of free parameters: $\tilde{w}_i = \tilde{d} + 1 \approx d_{vc}(\mathcal{H}_{\phi_Q})$
        - Since $\tilde{d} + 1$ dimensions is required to learn a linear model in $\mathcal{Z}$ space, any $\tilde{d} + 2$ inputs **cannot be shattered** in $\mathcal{Z}$-space
            - By definition, this implies $d_{vc}(\mathcal{H}_{\phi_Q}) \leq \tilde{d} + 1$
            - No line in $\mathcal{Z}$-space shatters $\tilde{d} + 2$ inputs $\Rightarrow$ No line in χ-space shatters $\tilde{d} + 2$ inputs (in χ-space)
    - **Large Q** $\Rightarrow$ Large d_{vc}
3. Higher-order nonlinear transformation also risk **overfitting** the training data set
    - **Generalization issue**



Φ_1 (original **x**) Φ_4

4. Keep in mind the two objectives of learning (and tradeoff between the two)
    - Make $E_{out}(g)$ **close** enough to $E_{in}(g)$
    - Make $E_{in}(g)$ **small** enough

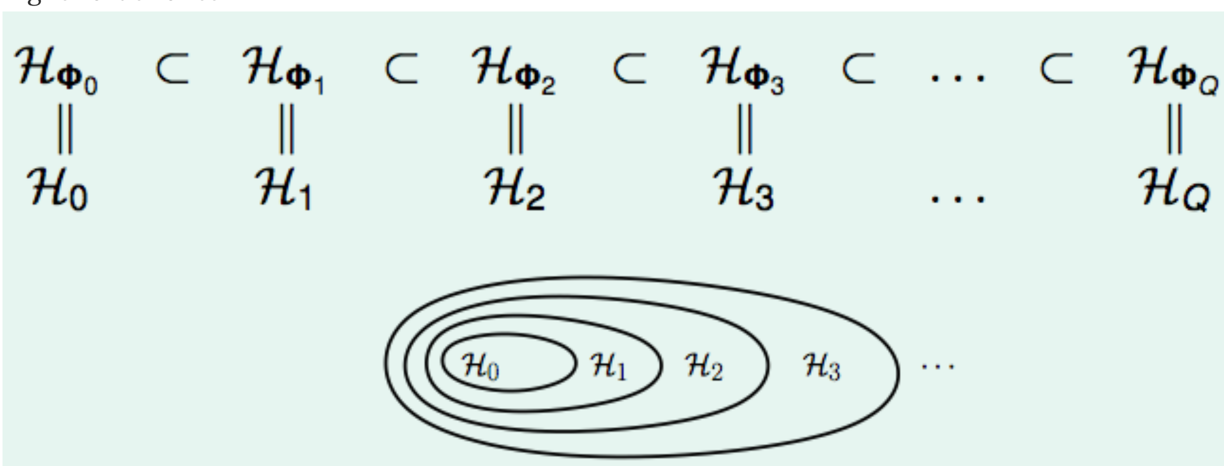
$\tilde{d}$(Q)	E_{in} close to E_{out}	Small E_{in}
High	Bad	Good
Low	Good	Bad

Structured Hypothesis Set

1. Given polynomial transforms from 0-th order to Q-th order:

$$
\begin{aligned}
\phi_0(x) &= 1 \\
\phi_1(x) &= (\phi_0(x), x_1, x_2, \ldots, x_d) \\
\phi_2(x) &= (\phi_1(x), x_1^2, x_1x_2, \ldots, x_d^2) \\
\phi_3(x) &= (\phi_2(x), x_1^3, x_1^2x_2, \ldots, x_d^3) \\
&\ldots\ldots \\
\phi_Q(x) &= (\phi_{Q-1}(x), x_1^Q, x_1^{Q-1}x_2, \ldots, x_d^Q)
\end{aligned}
$$

   - In other words, there is a **nested** relationship from lower-order polynomial transformations to higher order ones

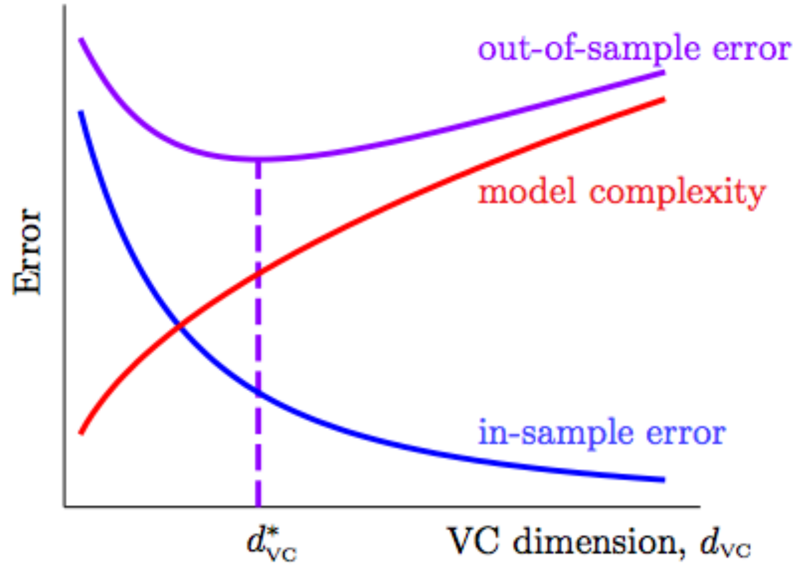


2. Model complexity is positively correlated with order of nonlinear transform. In-sample error negatively correlates with order of nonlinear transform. For optimal model learned with i-th order nonlinear transform, the following inequalities hold.

Let $g_i = \text{argmin}_{h \in \mathcal{H}_i} E_{\text{in}}(h)$:

$$
\begin{array}{ccccccccc}
\mathcal{H}_0 & \subset & \mathcal{H}_1 & \subset & \mathcal{H}_2 & \subset & \mathcal{H}_3 & \subset & \ldots \\
d_{\text{vc}}(\mathcal{H}_0) & \leq & d_{\text{vc}}(\mathcal{H}_1) & \leq & d_{\text{vc}}(\mathcal{H}_2) & \leq & d_{\text{vc}}(\mathcal{H}_3) & \leq & \ldots \\
E_{\text{in}}(g_0) & \geq & E_{\text{in}}(g_1) & \geq & E_{\text{in}}(g_2) & \geq & E_{\text{in}}(g_3) & \geq & \ldots
\end{array}
$$

3. However, as model complexity grows, there is a **decreasing marginal return** in terms of in-sample error reduction. And out-of-sample error tends to be large with complex models since they don't generalize well.



4. Rule of thumb for choosing models: **Linear model first**
    - Start from $\mathcal{H}_1$, because linear models are simple, efficient, **safe** (from generalization problems), and often produe workable results for the problem of interest
    - If $\mathcal{H}_1$ fails to produce plausible result, move right on model complexity curve. Computation wasted on $\mathcal{H}_1$ is still minimal