Machine Learning from Data

Lecture 13: Spring 2021

Today’s Lecture

- Validation and Model Selection
  - Validation Set
  - Model Selection
  - Cross validation

Regularization (Recap)

Regularization combats the effects of noise by putting a leash on the algorithm.

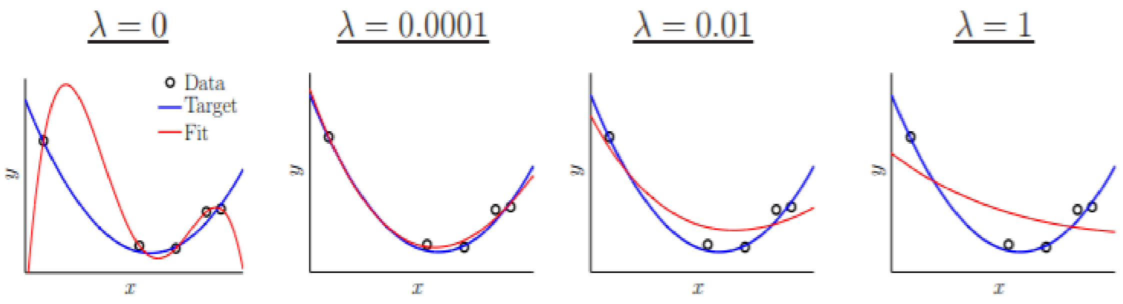
$$E_{\text{aug}}(h) = E_{\text{in}}(h) + \frac{\lambda}{N}\Omega(h)$$

$\Omega(h) \to$ smooth, simple h

noise is rough, complex.

Different regularizers give different results

can choose λ, the **amount** of regularization.



Overfitting $\to$ $\to$ **Underfitting**

Optimal λ balances approximation and generalization, bias and variance.

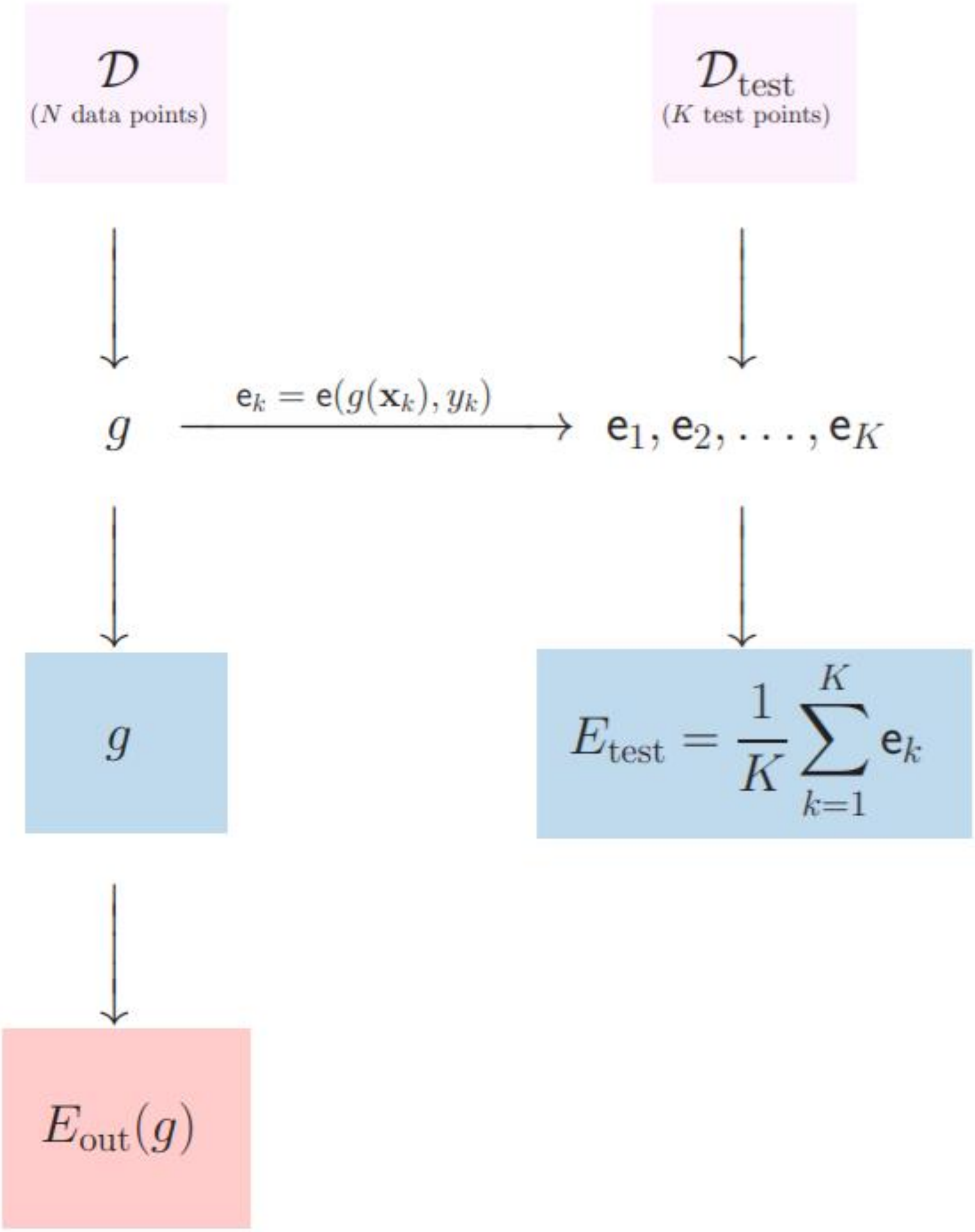
Validation

$$E_{\text{out}}(g) = E_{\text{in}}(g) + \underbrace{\text{overfit penalty}}_{\substack{\text{VC bounds this using a complexity error bar } \Omega(\mathcal{H}) \\ \text{regularization estimates this through a heuristic complexity penalty } \Omega(g)}}$$

Validation goes directly for the jugular:

$$\underbrace{E_{\text{out}}(g)}_{\text{validation estimates this directly}} = E_{\text{in}}(g) + \text{overfit penalty}.$$

In-sample estimate of E_{out} is the Holy Grail of learning from data.



E_{test} is an estimate for $E_{\text{out}}(g)$

$$\mathbb{E}_{\mathcal{D}_{\text{test}}}[\mathsf{e}_k] = E_{\text{out}}(g)$$

$$\begin{aligned}\mathbb{E}[E_{\text{test}}] &= \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\mathsf{e}_k] \\ &= \frac{1}{K}\sum_{k=1}^{K} E_{\text{out}}(g) = E_{\text{out}}(g)\end{aligned}$$

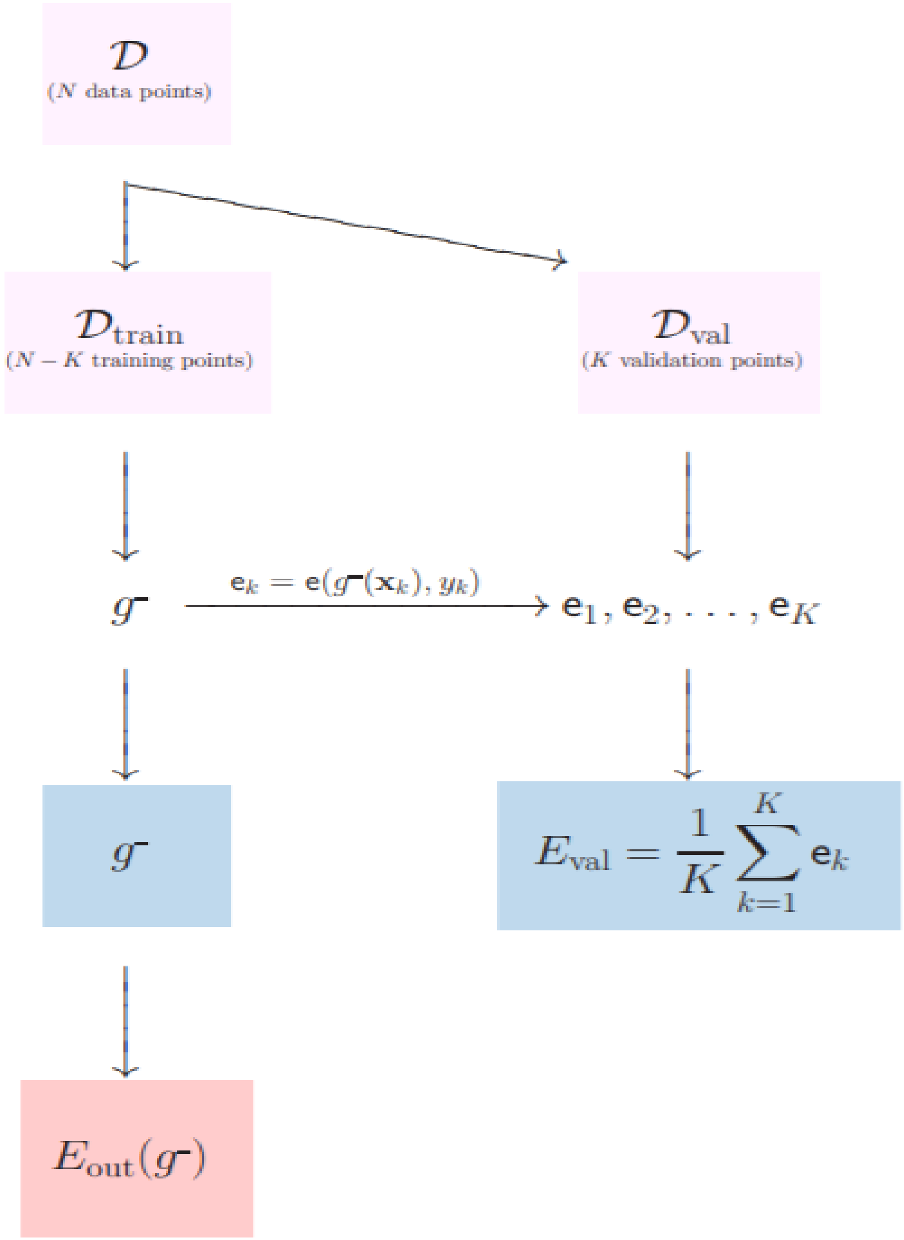
$\mathsf{e}_1, \ldots, \mathsf{e}_K$ are *independent*

$$\begin{aligned}\text{Var}[E_{\text{test}}] &= \frac{1}{K^2}\sum_{k=1}^{K} \text{Var}[\mathsf{e}_k] \\ &= \frac{1}{K}\text{Var}[e]\end{aligned}$$

decreases like $\frac{1}{K}$

bigger $K \implies$ more reliable E_{test}.

The Validation Set



E_{val} is an estimate for $E_{\text{out}}(g^{-})$

$$\mathbb{E}_{\mathcal{D}_{\text{val}}}[\mathsf{e}_k] = E_{\text{out}}(g^{-})$$

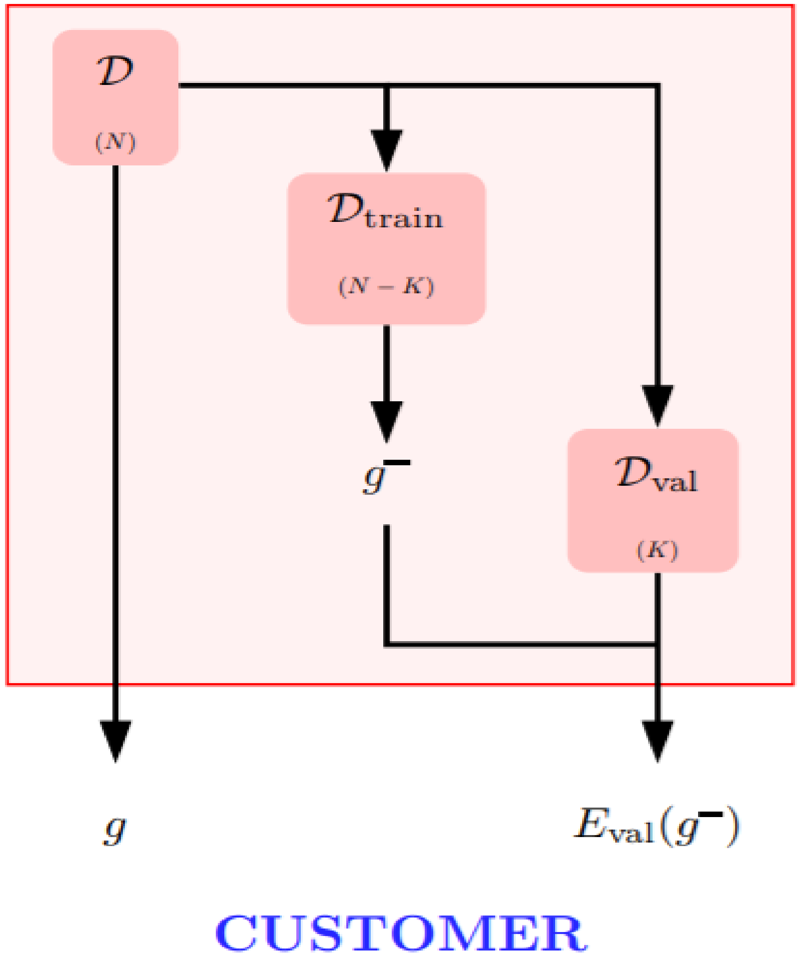
$$\begin{aligned}\mathbb{E}[E_{\text{test}}] &= \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}[\mathsf{e}_k] \\ &= \frac{1}{K}\sum_{k=1}^{K} E_{\text{out}}(g^{-}) = E_{\text{out}}(g^{-})\end{aligned}$$

$\mathsf{e}_1, \ldots, \mathsf{e}_K$ are *independent*

$$\begin{aligned}\text{Var}[E_{\text{val}}] &= \frac{1}{K^2}\sum_{k=1}^{K} \text{Var}[\mathsf{e}_k] \\ &= \frac{1}{K}\text{Var}[e(g^{-})]\end{aligned}$$

decreases like $\frac{1}{K}$
depends on g^{-}, not $\mathcal{H}$
bigger $K \implies$ more reliable E_{val}?

Restoring $\mathcal{D}$



Primary goal: output best hypothesis.

g was trained on *all* the data.

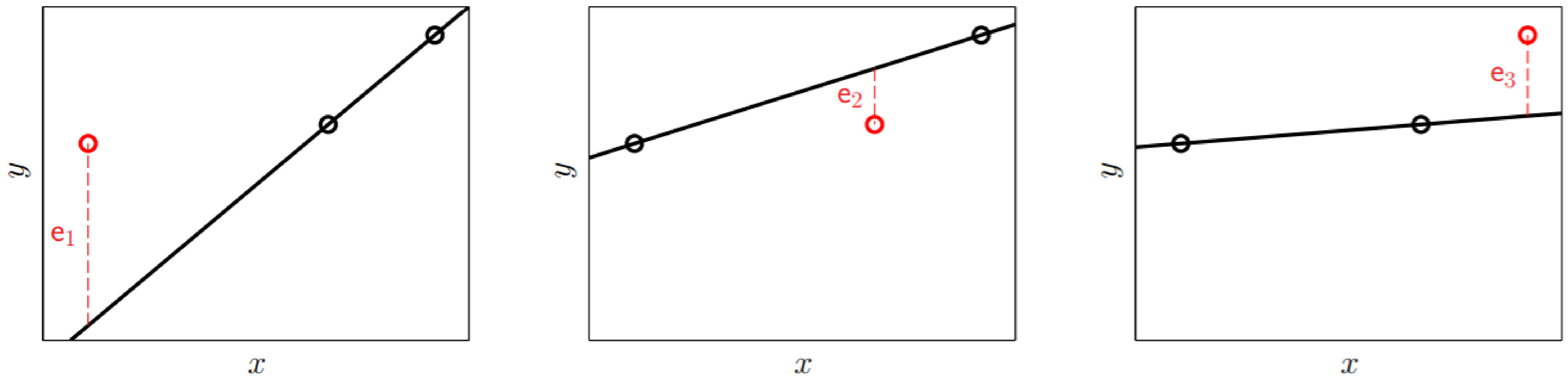
Secondary goal: estimate $E_{\text{out}}(g)$.

g^{-} is behind closed doors.

$$\underbrace{\begin{array}{cc} E_{\text{out}}(g) & E_{\text{out}}(g^{-}) \\ \downarrow & \downarrow \\ E_{\text{in}}(g) & E_{\text{val}}(g^{-}) \end{array}}_{\text{which should we use?}}$$

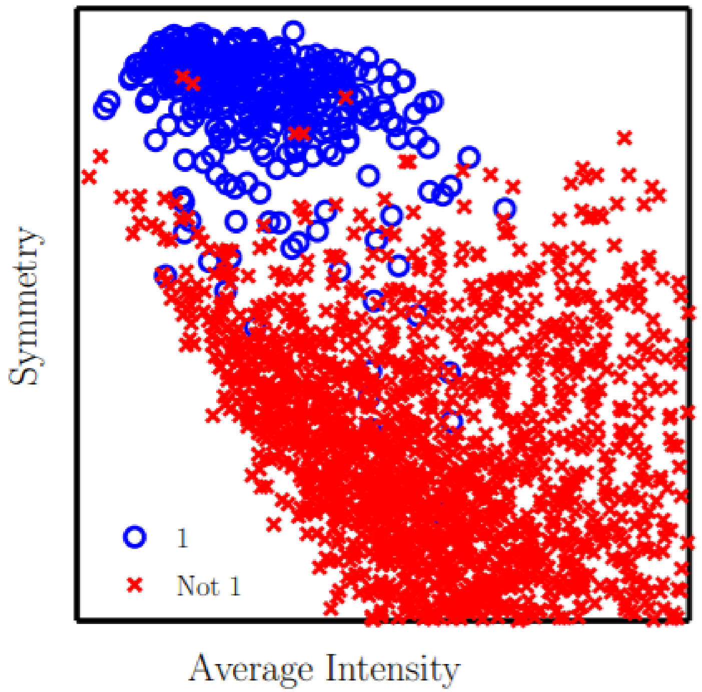
Model Selection

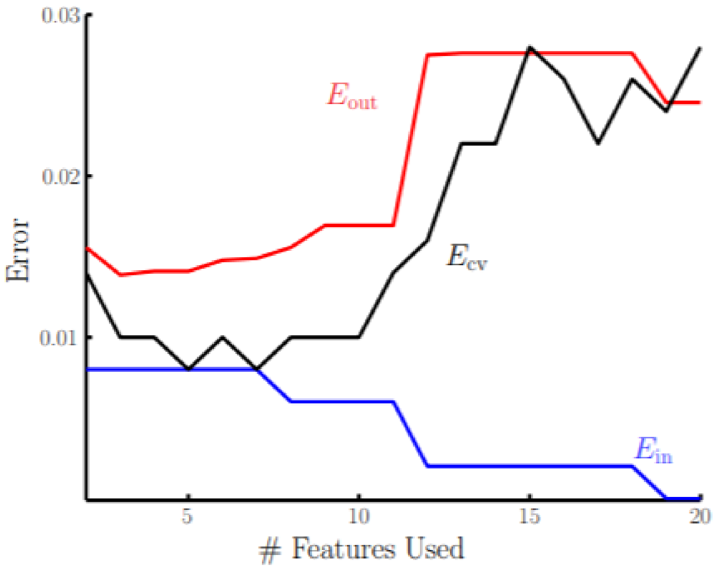
-



$$E_{\mathrm{cv}} = \frac{1}{N}\sum_{n=1}^{N} \mathsf{e}_n$$

Digits Problem: ‘1’ Versus ‘Not 1’

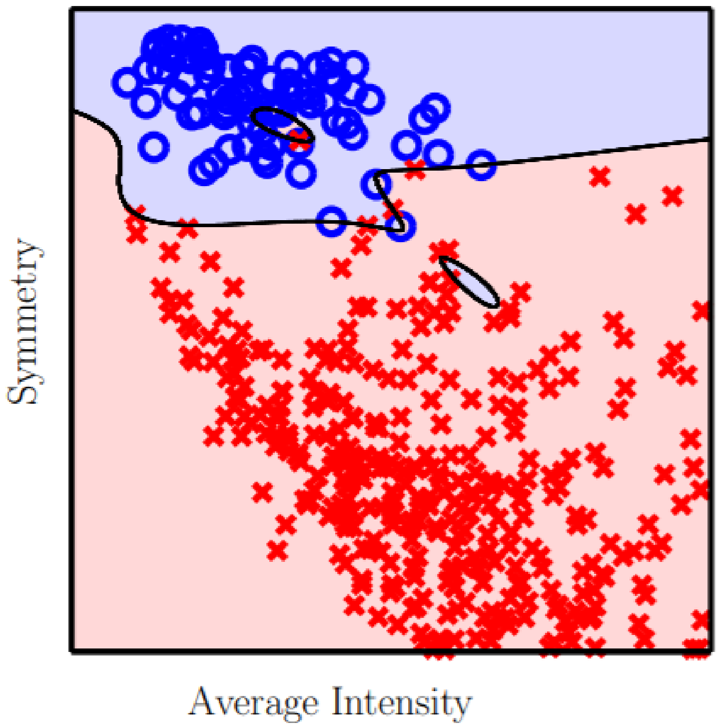




$$\mathbf{x} = (1, x_1, x_2)$$

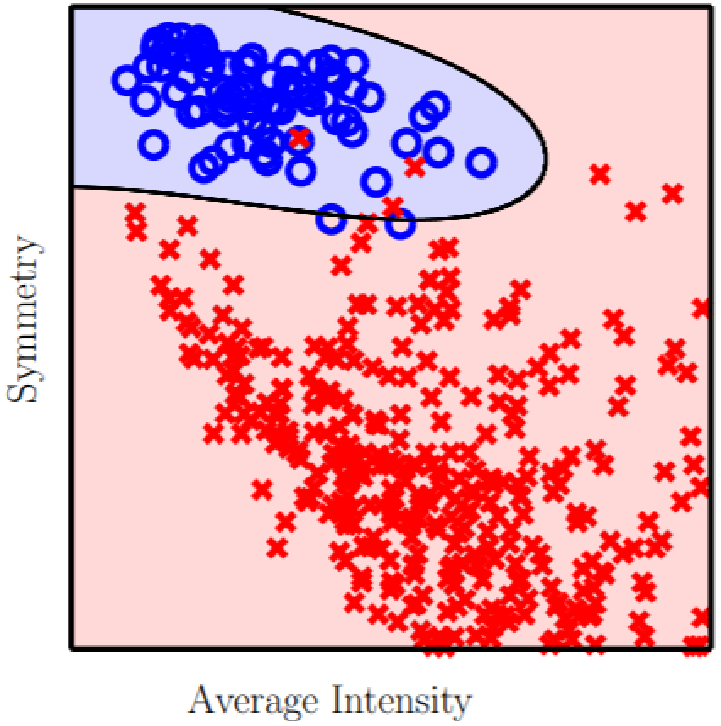
$$\mathbf{z} = \underbrace{(1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1^3, x_1^2x_2, x_1x_2^2, x_2^3, \ldots, x_1^5, x_1^4x_2, x_1^3x_2^2, x_1^2x_2^3, x_1x_2^4, x_2^5)}_{\text{5th order polynomial transform} \longrightarrow \text{20 dimensional non linear feature space}}$$

Validation Wins In the Real World



no validation (20 features)

$E_{\text{in}} = 0\%$
$E_{\text{out}} = 2.5\%$



cross validation (6 features)

$E_{\text{in}} = 0.8\%$
$E_{\text{out}} = \mathbf{1.5\%}$

Thanks!