

Complexity, Overfitting, and Cross-Validation

Objectives

- Create complex decision boundaries from measured data
- Examine complexity/performance tradeoffs
 - noise and errors
 - limited training data
- Introduce cross validation for performance evaluation

Add functions of data to generalize decision boundary²



Happy: +



Sad -

$$\textcircled{1} \quad \underline{x}^T = [x_1 \ x_2]$$

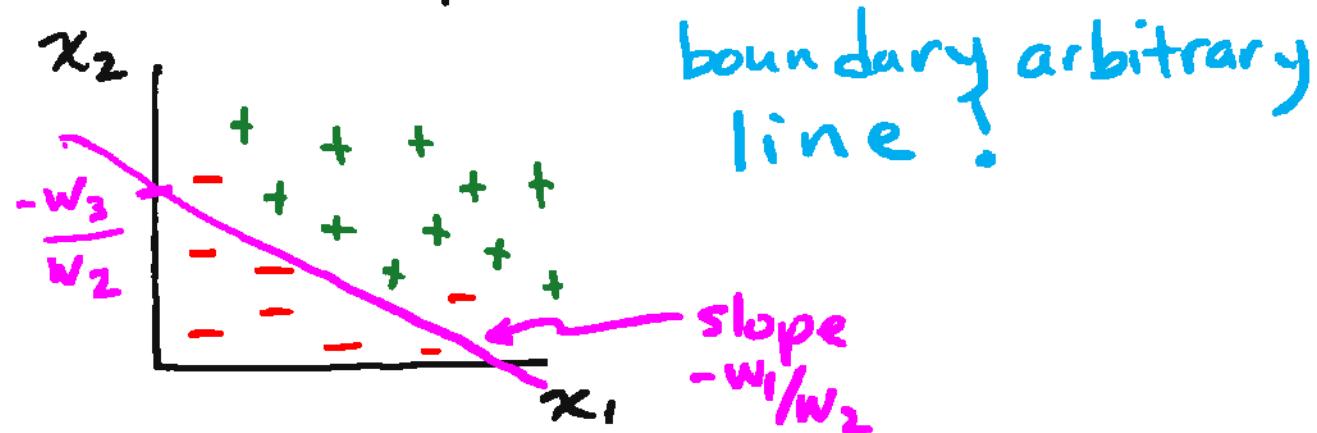
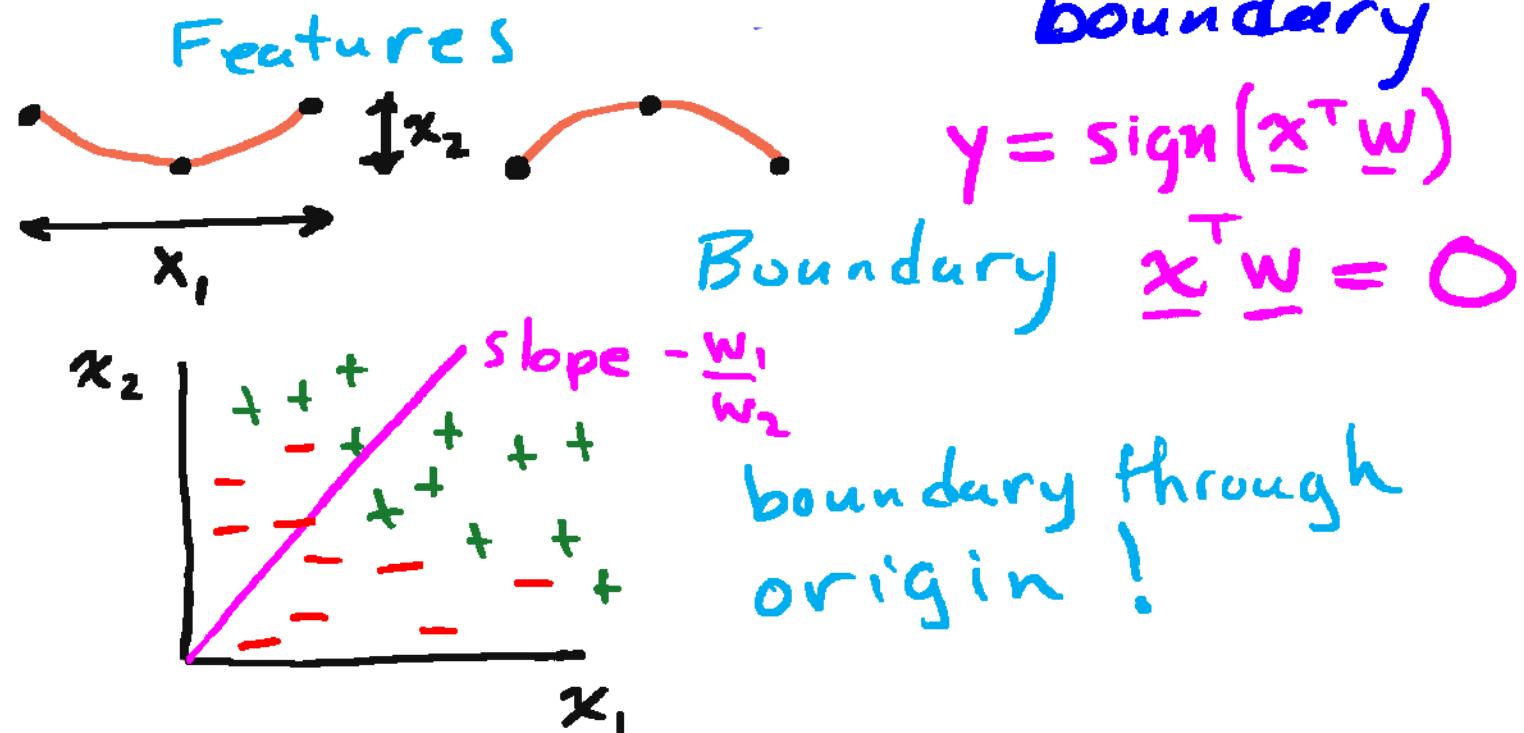
$$\text{Boundary } x_1 w_1 + x_2 w_2 = 0$$

$$x_2 = -\frac{w_1}{w_2} x_1$$

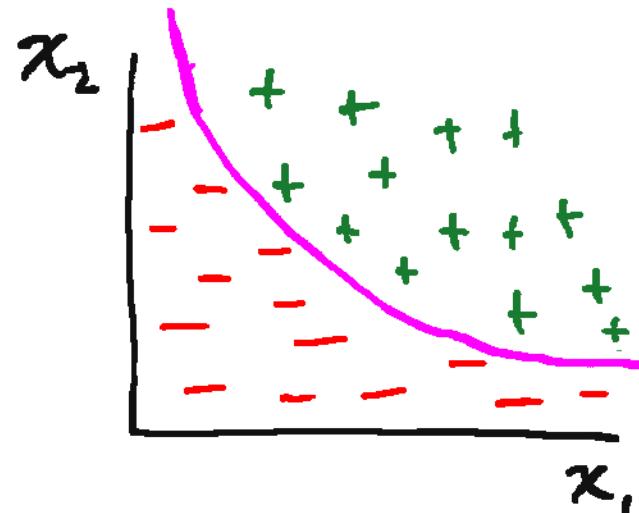
$$\textcircled{2} \quad \underline{x}^T = [x_1 \ x_2 \ 1]$$

$$\text{Boundary } x_1 w_1 + x_2 w_2 + w_3 = 0$$

$$x_2 = -\frac{w_1}{w_2} x_1 - \frac{w_3}{w_2}$$



What if we want a nonlinear boundary? 3



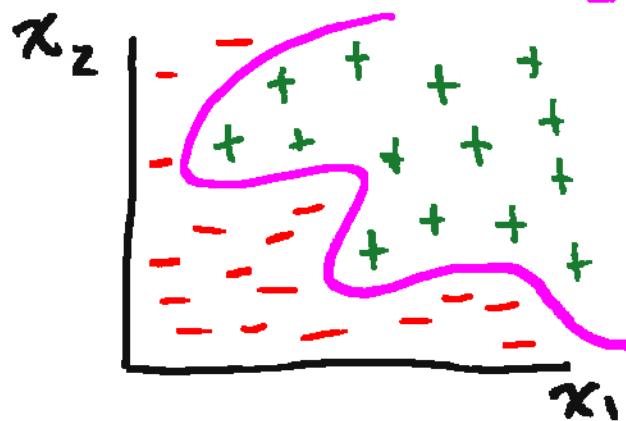
Add functions of x_1, x_2 to \underline{x}

③ $\underline{x}^T = [x_1 \ x_2 \ x_1^2 \ 1] \Rightarrow \underline{x}^T \underline{w} = 0$

$$x_2 = -\frac{w_3}{w_2} x_1^2 - \frac{w_1}{w_2} x_1 - \frac{w_4}{w_2}$$

parabolic boundary

In general $\underline{x}^T = [x_1 \ x_2 \ 1 \ x_1 x_2 \ x_1^2 \ x_2^2 \ x_1^2 x_2 \dots]$
arbitrarily complex boundary



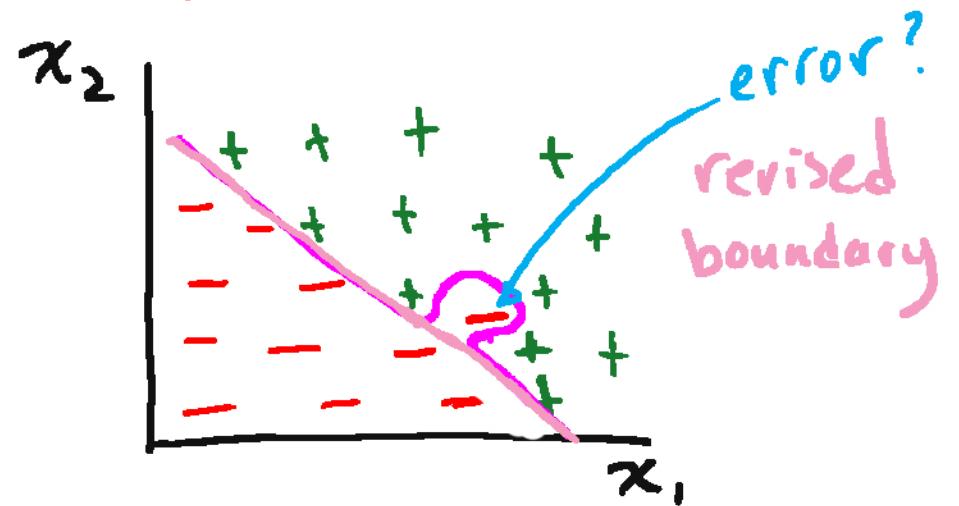
Includes simpler boundaries too!

Too good to be true?

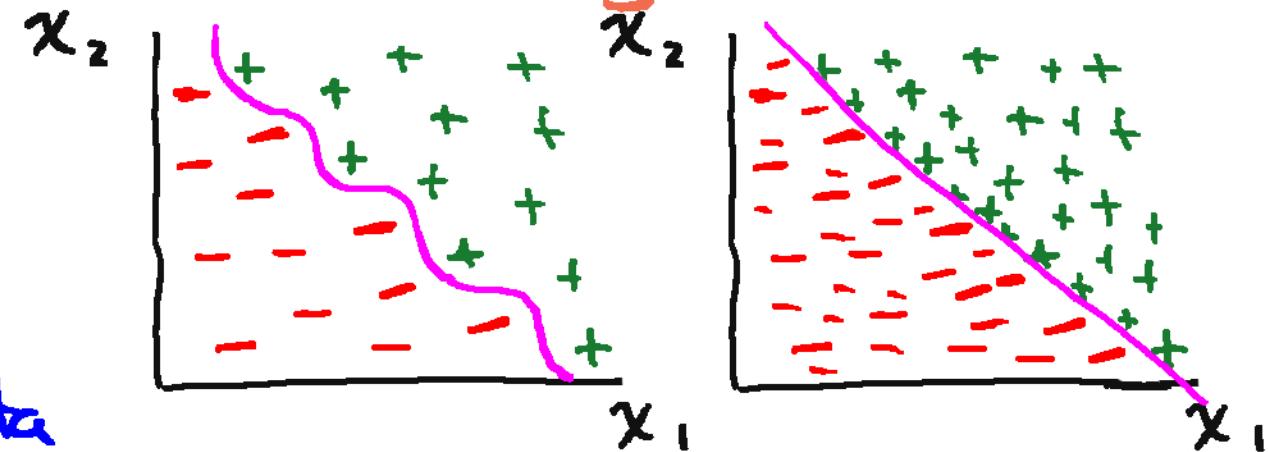
Number of model parameters ↑

Training data places limits on boundary complexity⁴

1. Presence of noise/errors



2. Limited quantity of training data



Goal: Generalize to new data

Over fitting - using too complex a model for training data

$P = \dim(\underline{w}) \uparrow$, training error \downarrow

In general, $P \ll N$

Cross validation measures generalization performance 5

- use subset of data to learn the model
- use held out data to evaluate performance

Divide $\{\underline{x}_i, y_i, i=1, 2, \dots, N\}$ into M groups $S_j = \{\underline{x}_i, y_i, i \in Q_j\}, j=1, 2, \dots, M$

Learn \underline{w}_1 from S_2, S_3, \dots, S_M . Compute $E_1 = \frac{1}{|Q_1|} \sum_{i \in Q_1} \frac{|y_i - \text{sign}(\underline{x}_i^T \underline{w})|}{2}$

Learn \underline{w}_2 from S_1, S_3, \dots, S_M . Compute E_2 using \underline{w}_2, S_2
⋮

Learn \underline{w}_M from S_1, S_2, \dots, S_{M-1} . Compute E_M using \underline{w}_M, S_M
⋮

$$E = \frac{1}{M} \sum_{i=1}^M E_i \quad \text{Average CV error}$$

Cross validation for model comparison

- prefer model with smaller CV score
- manage model complexity for bias/variance trade off
- identical considerations apply to modeling of data. Squared error performance:
$$E_i = \frac{1}{|Q_i|} \sum_{i \in Q_i} (\gamma_i - \underline{x}_i^T \underline{w})^2$$

Copyright 2019
Barry Van Veen