CSC 461: Machine Learning
Fall 2024

Regularization

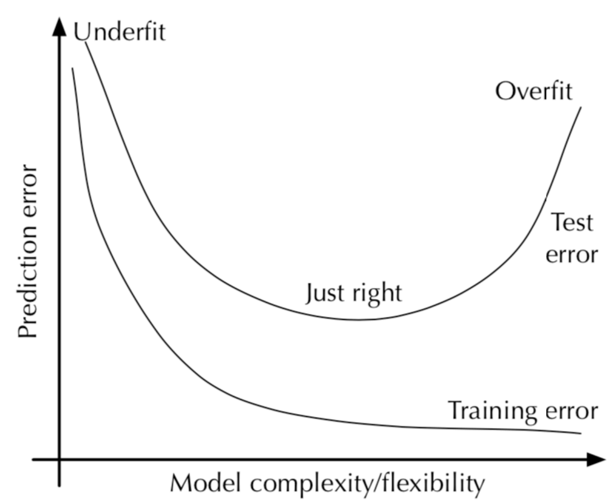
Prof. Marco Alvarez, Computer Science
University of Rhode Island

Overfitting

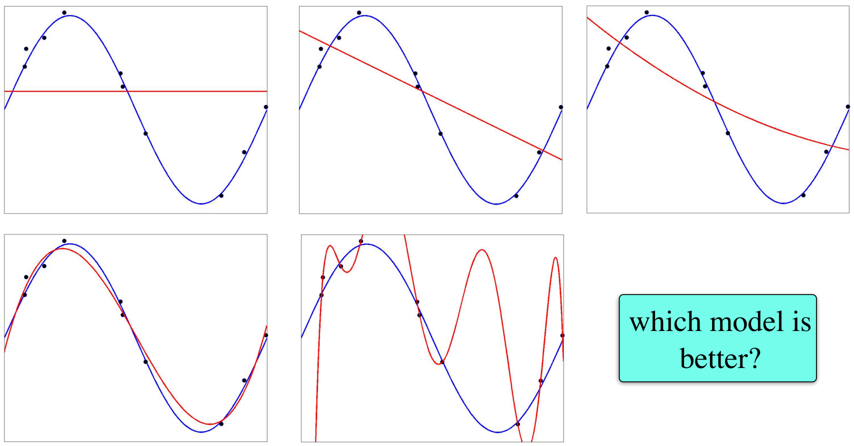
Model complexity and overfitting

- Manifestations of overfitting
  - complex model captures noise in training data
  - poor generalization to unseen data
  - high variance in predictions across different training sets
- How to prevent?
  - use more training data
  - use fewer features
  - **regularize** your model

Model complexity



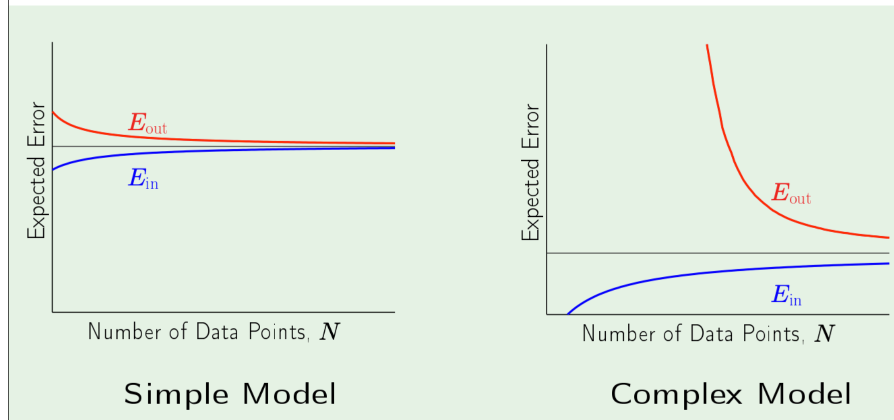
Overfitting and model complexity



which model is better?

Underfitting: model is too simple
Overfitting: model is too complex

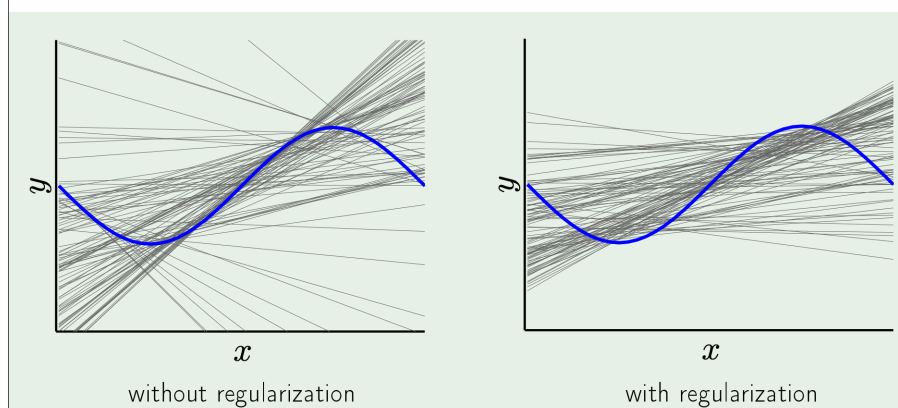
Overfitting and data size



https://work.caltech.edu/lectures.html

Regularization

Restricting the hypothesis space



https://work.caltech.edu/lectures.html

Regularization

- Original objective

$$\arg\min_{\mathbf{w}} \; L(\mathbf{w})$$

- Regularized objective

$$\arg\min_{\mathbf{w}} \; L(\mathbf{w}) + \lambda R(\mathbf{w})$$

- Common regularization terms
  - L1, L2, elastic net

Linear regression and regularization

- Control the complexity of the model
  - usually **penalizing higher weights** (except intercept)
  - results in simpler or more sparse solutions
- Impact of regularization can be controlled by a parameter (lambda)

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda R(\mathbf{w})$$

Linear regression and regularization

- L2 regularization
  - **Ridge Regression**
  - closed-form solution exists
    - $(X^T X + \lambda I)^{-1} X^T \mathbf{y}$
  - differentiable everywhere
  - shrinks all weights proportionally

$$R(\mathbf{w}) = \|\mathbf{w}\|_2^2$$

- L1 regularization
  - **Lasso Regression**
  - does not have a closed-form solution (not differentiable)
  - promotes sparsity

$$R(\mathbf{w}) = \|\mathbf{w}\|_1$$

Linear regression and regularization

$$L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{w}^T\mathbf{x}^{(i)} - y^{(i)}\right)^2 + \lambda \sum_{k=1}^{d} w_k^2$$

partial derivatives with respect to a single w_j

$$\frac{\partial L(\mathbf{w})}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n}\left(\mathbf{w}^T\mathbf{x}^{(i)} - y^{(i)}\right)$$

don't regularize the intercept

$$\frac{\partial L(\mathbf{w})}{\partial w_j} = \frac{2}{n}\sum_{i=1}^{n}\left(\mathbf{w}^T\mathbf{x}^{(i)} - y^{(i)}\right)x_j^{(i)} + 2\lambda w_j$$

Practice

$$\mathbf{x} = [1,1,1,1] \qquad \mathbf{w_a} = [1,0,0,0]$$

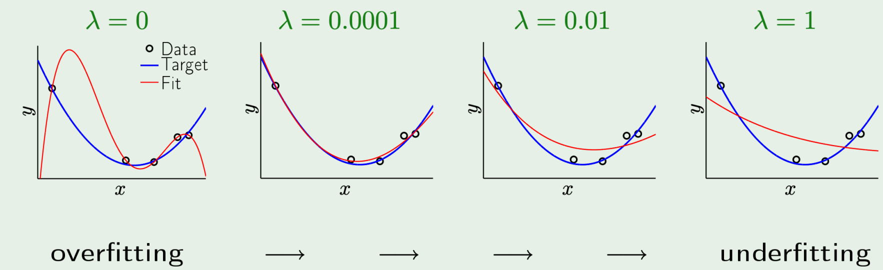
$$\mathbf{w_b} = [.25,.25,.25,.25]$$

Assume linear regression, what is $h(\mathbf{x})$ for each solution $\mathbf{w_a}$ and $\mathbf{w_b}$?

Which of the solutions will the L2 regularizer prefer?

Which of the solutions will the L1 regularizer prefer?

Regularization strength



https://work.caltech.edu/lectures.html