Overfitting

Author: Ziqi Tan

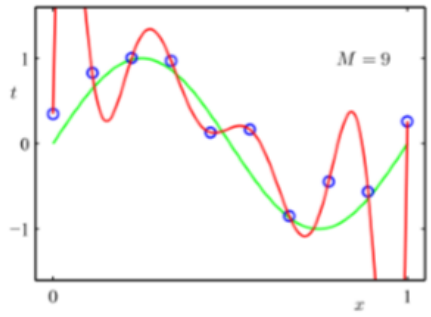
Date: Feb 23, 2020

Consequence

Parameters for higher-order polynomials are very large

	M =0	M =1	M =3	M =9
θ_0	0.19	0.82	0.31	0.35
θ_1		-1.27	7.99	232.37
θ_2			-25.43	-5321.83
θ_3			17.37	48568.31
θ_4				-231639.30
θ_5				640042.26
θ_6				-1061800.52
θ_7				1042400.18
θ_8				-557682.99
θ_9				125201.43

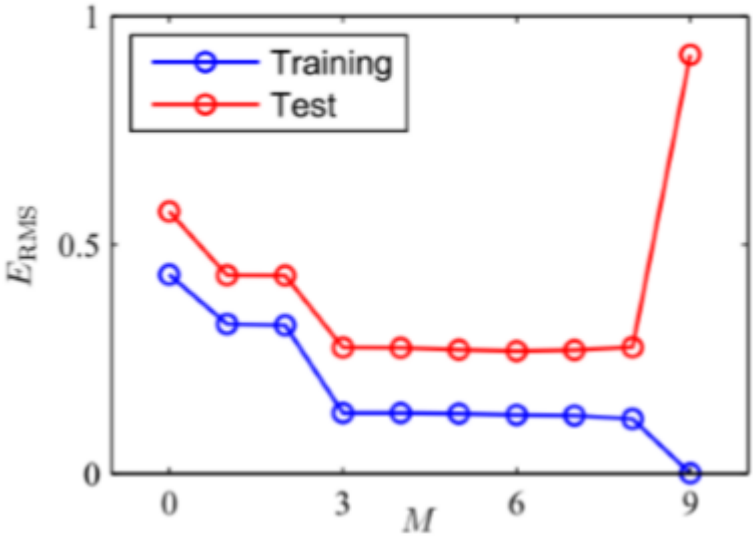
M=9: *overfitting*



Overfitting is also called **poor generalization**.

Detecting overfitting

Plot model complexity versus object function on test/train data.



As model becomes more complex, performance on training keeps improving while on test data it increases.

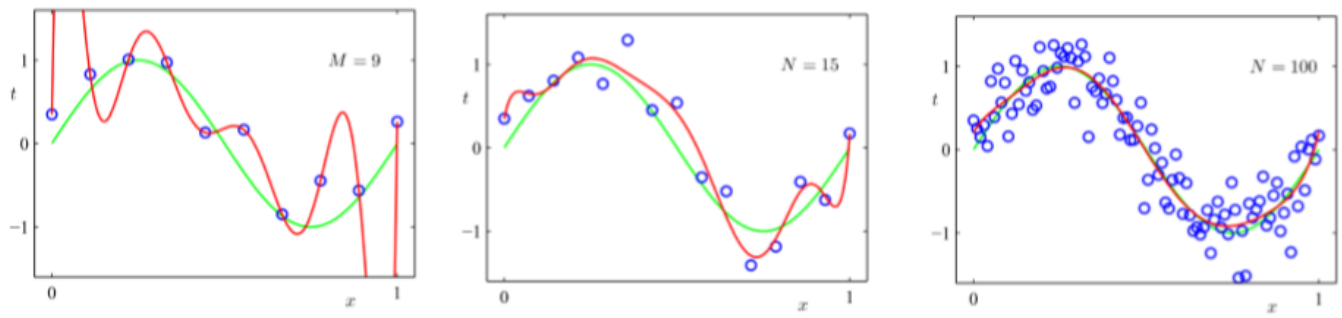
x-axis: measure of model complexity In this example, we use the maximum order of the polynomial basis functions.

y-axis:
For regression, it would be SSE or mean SE (MSE).
For classification, the vertical axis would be classification error rate or cross-entropy error function

Overcome overfitting

1. Use more training data.
2. Regularization.
3. Cross-validation.

M=9, increase N

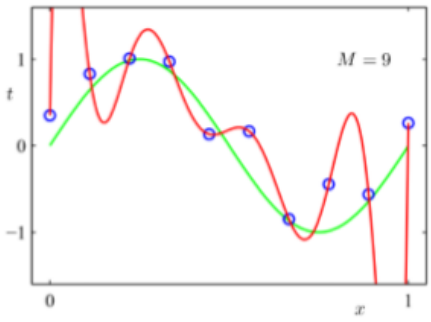


What if we do not have a lot of data?

Regularization

Solution: Regularization

- Use regularization:
  - Add $\lambda \|\theta\|_2^2$ term to SSE cost function
  - "L-2" norm squared, ie sum of sq. elements $\sum \theta_j^2$
  - Penalizes large θ
  - λ controls amount of regularization



	M =9
θ_0	0.35
θ_1	232.37
θ_2	-5321.83
θ_3	48568.31
θ_4	-231639.30
θ_5	640042.26
θ_6	-1061800.52
θ_7	1042400.18
θ_8	-557682.99
θ_9	125201.43

$$J(\theta) = \frac{1}{2m}[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2]$$

where the regularization can be written as a matrix form:

$$\theta^T A \theta$$

where A is a symmetric matrix, serving the same purpose as λ does.

L2 Regularization

1 norm:

$$||W||_1 = |W_1| + |W_2| + \ldots + |W_N|$$

2 norm:

$$||W||_2 = (|W_1|^2 + |W_2|^2 + \ldots + |W_N|^2)^{\frac{1}{2}}$$

p norm:

$$||W||_p = (|W_1|^p + |W_2|^p + \ldots + |W_N|^p)^{\frac{1}{p}}$$

Gradient descent

Repeat {

$$\theta_j = \theta_j - \alpha \frac{1}{m} [\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j + \lambda \theta_j]$$

} until convergence.

Cross-Validation

Validation用来调参和跑分，Test用来跑分，不用来调参。

Train/Validation/Test Sets

	Size	Price
train	2104	400
	1600	330
	2400	369
	1416	232
	3000	540
validation	1985	300
	1534	315
	1427	199
test	1380	212
	1494	243

Solution: split data into three sets.

For each value of a hyperparameter, train on the train set, evaluate learned parameters on the validation set.

Pick the model with the hyper parameter that achieved the lowest validation error.

Report this model's test set error.

N-Fold Cross Validation

- What if we don't have enough data for train/test/validation sets?
- Solution: use N-fold cross validation.
- Split training set into train/validation sets N times.
- Report average predictions over N val sets, e.g. N=10: