Topic 2 Model Selection continued

PROF. LINDA SELLIE

Outline

- ❑Motivating example:
- ❑ Feature transformation
- ❑Underfitting and overfitting
- ❑Understanding error: Bias and variance and noise
- ❑Learning curves
- ❑validation and model selection
- ❑Model selection (with limited data)
- ❑K-fold cross validation
- ❑Regularization

How to create a more complex hypothesis

Understanding where the error comes from, and how to estimate $E_{\text{out}}[g(\mathbf{x})]$

If we have many different hypothesis classes to choose from - how can we choose wisely? And how can we estimate $E_{\text{out}}[g(\mathbf{x})]$?

Outline

- ❑Motivating example:
- ❑ Feature transformation
- ❑Underfitting and overfitting
- ❑Understanding error: Bias and varianc
- ❑Learning curves
- ❑validation and model selection
- ❑Model selection (with limit
- ❑K-fold cross validation
- ❑Regularization

Yea!

Uh oh....

How to create a more complex hypothesis

Understanding what went wrong

Understanding where the error comes from, and how to estimate $E_{\text{out}}[g(\mathbf{x})]$

Our strategy

If we have many different hypothesis classes to choose from - how can we choose wisely? And how can we estimate $E_{\text{out}}[g(\mathbf{x})]$?

How do we evaluate our model? Or choose among models (e.g. the which polynomial transformation should we choose?)

Open discussion

- We can evaluate how well it works by looking at its errors
- We would like the error to be zero on all future data. However:
  - The unseen variables means the true model has non-zero error (i.e. the world is a messy place) — Noise/irreducible error
  - Our hypothesis probably doesn't contain the underlying true model — Bias
  - We don't get enough data to perfectly estimate our model. We only get a finite sample of the data. The more data we receive, the more our sample is representative of underlying data and our estimates should converge — Variance

Next: A Mathematical explanation of the error

You are not expected to do this in the homework.

You will not use these equations to determine the error of your model.

This is purely theoretical.

Where did the prediction error in our hypothesis come from?

- Regression example: $y = f(\mathbf{x}) + \epsilon$

Deterministic

Noise $\sim N(0,\sigma)$
We are assuming the noise has mean 0 and variance σ^2

This means $E_{\mathbf{x},y}[f(\mathbf{x}) - y] = 0$ and $E_{\mathbf{x},y}[(f(\mathbf{x}) - y)^2] = E_{\mathbf{x}}(\epsilon^2) = \sigma^2$

Best estimate for y given $\mathbf{x}$ is $f(\mathbf{x})$

- Goal is to understand why our *expected* hypothesis (model) does not have zero error

$$E_D\left[E_{\text{out}}(g^{(D)})\right] = E_D\left[E_{\mathbf{x},y}[(g^{(D)}(\mathbf{x}) - y)^2]\right] \neq 0$$

$E_{\text{out}}(g^{(D)})$

$E_{\mathbf{x},y}[(g^{(D)}(\mathbf{x}) - y)^2]$ expected error for they hypothesis $g^{(D)}(\mathbf{x})$

The expected error of the hypothesis on any future example. The hypothesis was fit using the data set D

We focus on Algorithm Bias

Understanding Error

Bias-Variance-Noise Decomposition

$$E_{\text{out}}(g^{(D)}(\mathbf{x})) = E_{\mathbf{x},y}[(y - g^{(D)}(\mathbf{x}))^2]$$

Our definitions will be for the squared loss function
You can think of how to substitute other loss functions

$$E_{\text{out}}(g) = \underline{\text{bias} + \text{variance} + \text{noise}}$$

This cannot be computed in practice because we do not have access to the target function or the probability distribution

In predictions there are three sources of error.

1. noise - irreducible error
2. bias - error of average hypothesis (estimated from N examples) from the true function
3. variance - how much would the prediction for an example change if the hypothesis was fit on a different set of N points

High Bias $\leftrightarrow$ underfitting

High Variance $\leftrightarrow$ overfitting

Outline

- Motivating example:
- Polynomial transformation
- Underfitting and overfitting
- Understanding error: Bias and variance and noise
  - Bias
  - Variance
  - Bias and variance and noise
- Learning curves
- validation
- Model selection
- Cross validation
- Regularization

Yea!

Uh oh….

How to create a more complex hypothesis

Understanding what went wrong

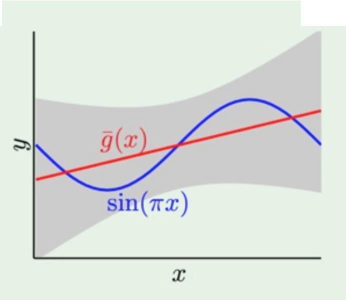
Understanding where the error comes from and how to estimate $E_{\text{out}}[g(\mathbf{x})]$

Our strategy

If we have many different hypothesis classes to choose from - how can we choose wisely? And how to estimate $E_{\text{out}}[g(\mathbf{x})]$?

Bias

Bias of the hypothesis class (not an individual hypothesis from the class)



- bias($\mathbf{x}$) = $\left(f(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2$

Conceptually: squared difference from "average prediction" for $\mathbf{x}$, and expected label $f(\mathbf{x})$

- bias $= E_{\mathbf{x}}\left[(\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2\right]$

$$\approx \frac{1}{N}\sum_{i=1}^{N}(\bar{g}(\mathbf{x}^{(i)}) - f(\mathbf{x}^{(i)}))^2$$

Bias of the hypothesis class

less flexible model then more bias

Occasionally this is called bias^2

When using this model class, measures how well you expect the "average prediction" to represent the true solution
We expect the bias to decreases with a more complex model

Outline

- Motivating example: What polynomial degree should a ?
- Polynomial transformation
- Underfitting and overfitting
- Understanding error: Bias and variance and noise
  - Bias
  - Variance
  - Bias and variance and noise
- Learning curves
- validation
- Model selection
- Cross validation
- Regularization

Yea!

Uh oh….

How to create a more complex hypothesis

Understanding what went wrong

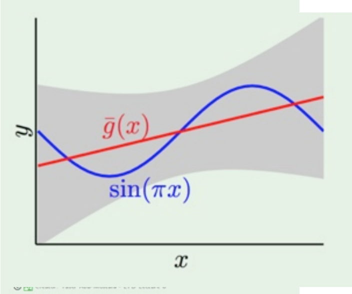
Understanding where the error comes from and how to estimate $E_{\text{out}}[g(\mathbf{x})]$

Our strategy

If we have many different hypothesis classes to choose from - how can we choose wisely? And how to estimate $E_{\text{out}}[g(\mathbf{x})]$?

Variance

Variance of a hypothesis class (model class)



- Variance: difference between the expected prediction and the prediction from a particular dataset

- $\mathrm{var}(\mathbf{x}) = E_D[(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2] \approx \frac{1}{L}\sum_{\ell=1}^{L}(\bar{g}(\mathbf{x}) - g_\ell^{(D_\ell)}(\mathbf{x}))^2$

Conceptually: variance of a prediction for $\mathbf{x}$ from the mean prediction

$$\mathrm{var} = E_{\mathbf{x}}\left[E_D\left[(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2\right]\right] \approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L}\sum_{\ell=1}^{L}(\bar{g}(\mathbf{x}^{(i)}) - g_\ell^{(D_\ell)}(\mathbf{x}^{(i)}))^2$$

less flexible model then less variance

Measures how sensitive a hypothesis class (model class) is to a specific dataset
Variance typically decreases with simpler models

Outline

- Motivating example:
- Feature transformation
- Underfitting and overfitting
- Understanding error: Bias and variance and noise
  - Bias
  - Variance
  - Bias and variance (and noise)
- Learning curves
- validation
- Model selection
- Cross validation
- Regularization

Yea!

Uh oh....

How to create a more complex hypothesis

Understanding where the error comes from and how to estimate $E_{\text{out}}[g(\mathbf{x})]$

Understanding what went wrong

Our strategy

If we have many different hypothesis classes to choose from - how can we choose wisely? And how to estimate $E_{\text{out}}[g(\mathbf{x})]$?

Generalization error: bias, variance, noise decomposition

The expected error of the hypothesis $g^{(D)}(\mathbf{x})$ on any future example. The model was fit using the data set D

$$E_{\text{out}}(g^{(D)}) = E_{\mathbf{x},y}[(g^{(D)}(\mathbf{x}) - y)^2]$$

The expected error of the hypothesis fit on a randomly chosen set of N training examples

$$E_D\left[E_{\text{out}}(g^{(D)})\right] = \underline{E_{\mathbf{x}}\left[E_D\left[\left(g^{(D)}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right]} + \sigma^2$$

$$E_{\mathbf{x}}\left[E_D[(g^{(D)}(\mathbf{x}) - f(\mathbf{x}))^2]\right]$$ = Bias + variance $\neq 0$

$y^{(i)} = f(\mathbf{x}^{(i)}) + \epsilon^{(i)}$ We are assuming the noise has mean 0 and variance σ^2

Understanding Error

$E_{\mathbf{x},y}[f(\mathbf{x}) - y] = 0 \quad E_{\mathbf{x},y}[(f(\mathbf{x}) - y)^2] = \sigma^2$

Bias-Variance-Noise Decomposition

The expected error of the hypothesis fit on a randomly chosen set of N training examples

$$E_D\left[E_{\text{out}}(g^{(D)})\right] = E_D\left[E_{\mathbf{x},y}[(g^{(D)}(\mathbf{x}) - y)^2]\right] = E_{\mathbf{x},y}\left[E_D[(g^{(D)}(\mathbf{x}) - y)^2]\right]$$

$$= E_{\mathbf{x},y}\left[E_D[(\underbrace{g^{(D)}(\mathbf{x}) - f(\mathbf{x})}_{A} + \underbrace{f(\mathbf{x}) - y}_{B})^2]\right]$$

$(A+B)^2=(A^2 + 2AB + B^2)$

Next: use linearity of expectation

$$= E_{\mathbf{x},y}\left[E_D[\left(g^{(D)}(\mathbf{x}) - f(\mathbf{x})\right)^2 + 2\left(g^{(D)}(\mathbf{x}) - f(\mathbf{x})\right)\left(f(\mathbf{x}) - y\right) + \left(f(\mathbf{x}) - y\right)^2]\right]$$

$$= E_{\mathbf{x},y}\left[E_D[\left(g^{(D)}(\mathbf{x}) - f(\mathbf{x})\right)^2] + 2E_D\left[\left(g^{(D)}(\mathbf{x}) - f(\mathbf{x})\right)\left(f(\mathbf{x}) - y\right)\right] + E_D[\left(f(\mathbf{x}) - y\right)^2]\right]$$

$$= E_{\mathbf{x},y}\left[E_D[\left(g^{(D)}(\mathbf{x}) - f(\mathbf{x})\right)^2] + 2E_D\left[\left(g^{(D)}(\mathbf{x}) - f(\mathbf{x})\right)\right]\underbrace{\left(f(\mathbf{x}) - y\right)}_{0} + \left(f(\mathbf{x}) - y\right)^2\right]$$

y doesn't occur here

$$= E_{\mathbf{x},y}\left[E_D[\left(g^{(D)}(\mathbf{x}) - f(\mathbf{x})\right)^2]\right] + \sigma^2$$

Understanding Error
Bias-Variance Decomposition (noise free)

$$\text{bias}(\mathbf{x}) = \left(f(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2 \qquad \text{var}(\mathbf{x}) = E_D[(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2]$$

$$\text{bias} = E_{\mathbf{x}}\left[(\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2\right] \qquad \text{var} = E_{\mathbf{x}}\left[E_D\left[(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2\right]\right]$$

$$\bar{g}(\mathbf{x}) = E_D[g^{(D)}(\mathbf{x})]$$

$$E_{\mathbf{x}}\left[E_D\left[\left(g^{(D)}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right] = E_{\mathbf{x}}\left[E_D\left[(\underbrace{g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x})}_{A} + \underbrace{\bar{g}(\mathbf{x}) - f(\mathbf{x})}_{B})^2\right]\right]$$

$$= E_{\mathbf{x}}\left[E_D\left[\left(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2 + 2\left(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right) + \left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right]$$

$$= E_{\mathbf{x}}\left[E_D\left[\left(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right] + 2E_D\left[\left(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)\right] + E_D\left[\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]\right]$$

$$= E_{\mathbf{x}}\Big[\underbrace{E_D\left[\left(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]}_{\text{variance}(\mathbf{x})} + 2\underbrace{E_D\left[\left(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)\right]}_{0}\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right) + \underbrace{E_D\left[\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]}_{\text{bias}(\mathbf{x})}\Big]$$

$$= E_{\mathbf{x}}[\text{bias}] + E_{\mathbf{x}}[\text{variance}]$$

$$= \text{bias} + \text{variance}$$

Notice that

$$E_D\left[\left(g^{(D)}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)\right]$$

$$= E_D[g^{(D)}(\mathbf{x})] - \bar{g}(\mathbf{x})$$

Understanding Error
Bias-Variance-Noise Decomposition

The expected error of the hypothesis fit on a randomly chosen set of N training examples

$$E_{\text{out}}(g) \;=\; \text{bias} + \text{variance} + \text{noise}$$

Noise in the training set contributes to variance

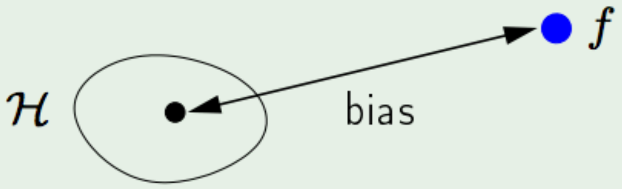
Noise in the test set contributes to irreducible error

Based on averages over what is expected for a training set D

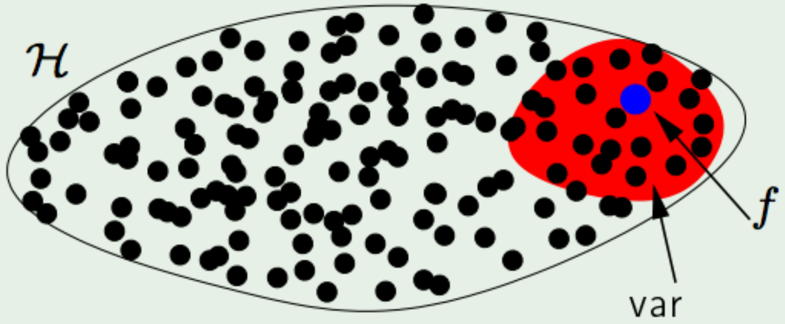
- can we lower variance without increasing too much the bias?
- can we lower bias without increasing too much the variance?

The tradeoff

$$\text{bias} = \mathbb{E}_{\mathbf{x}}\left[\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] \qquad \text{var} = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]\right]$$



bias $= E_{\mathbf{x}}$



$$\downarrow \qquad \mathcal{H} \uparrow \qquad \uparrow$$

Example: sine target

$f : [-1, 1] \to \mathbb{R} \qquad f(x) = \sin(\pi x)$

Only two training examples! $\quad N = 2$

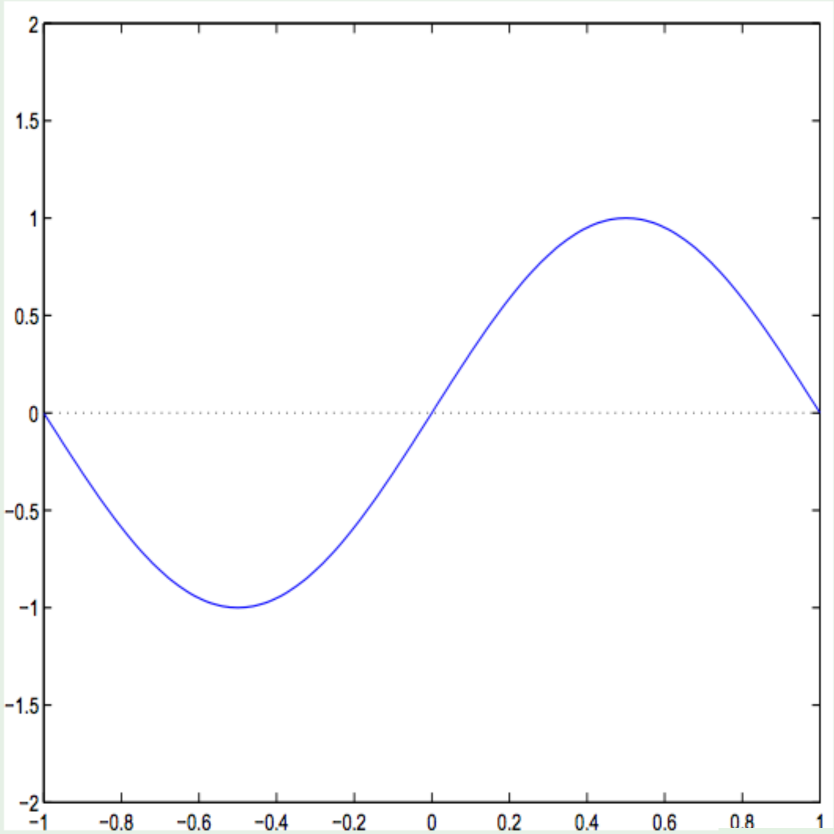
Two models used for learning:

$\mathcal{H}_0$: $\quad g(x) = w_0$

$\mathcal{H}_1$: $\quad g(x) = w_0 + w_1 x$

Which is better, $\mathcal{H}_0$ or $\mathcal{H}_1$?

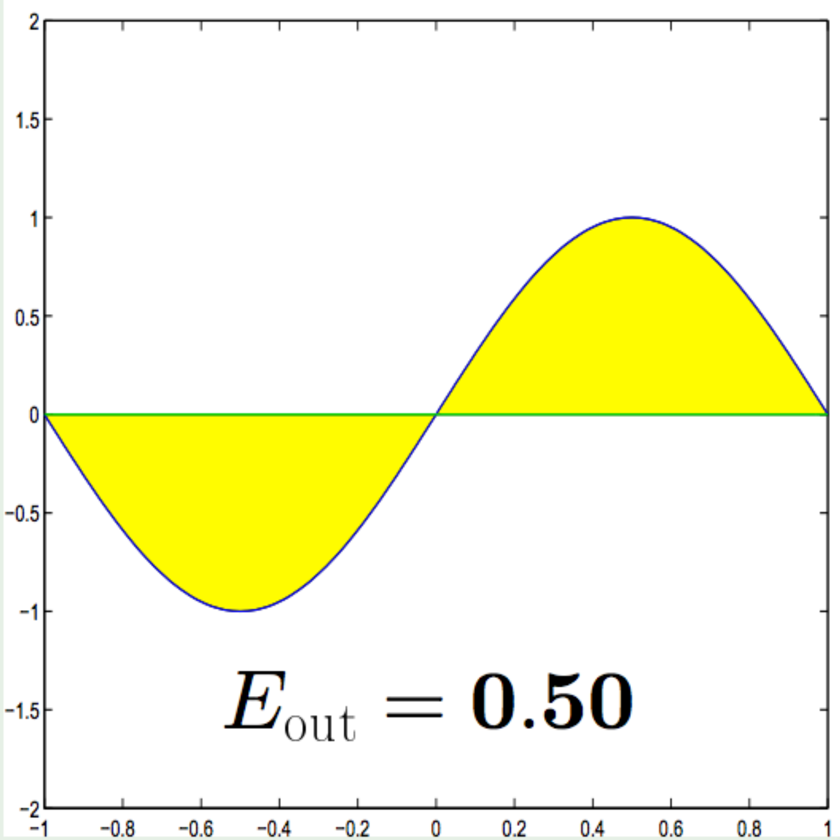
f



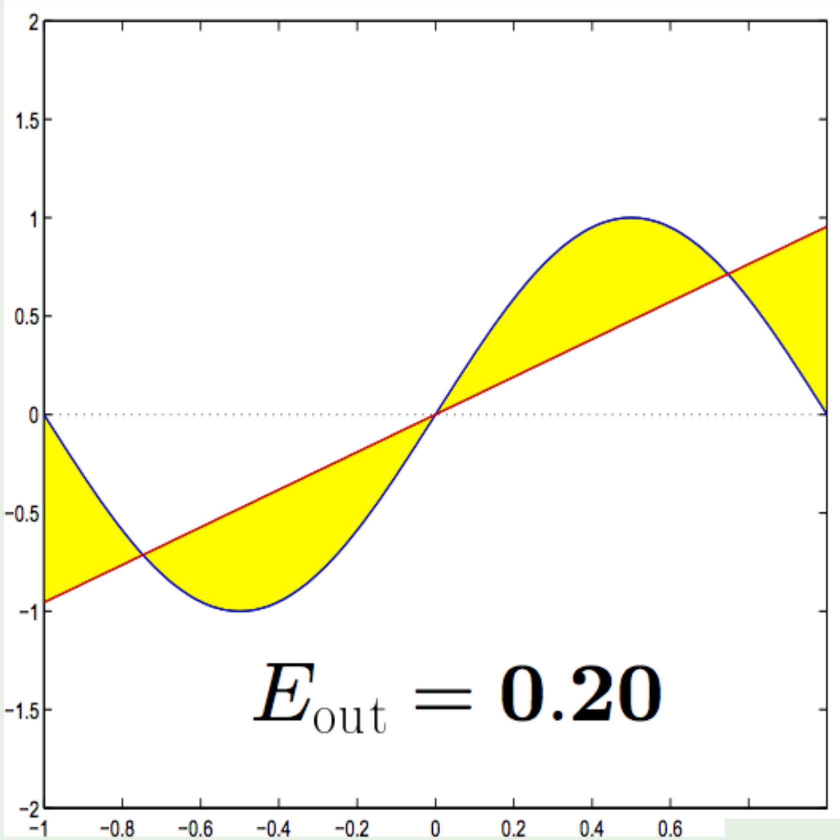
Very slightly modified

Approximation - $\mathcal{H}_0$ versus $\mathcal{H}_1$

$\mathcal{H}_0$

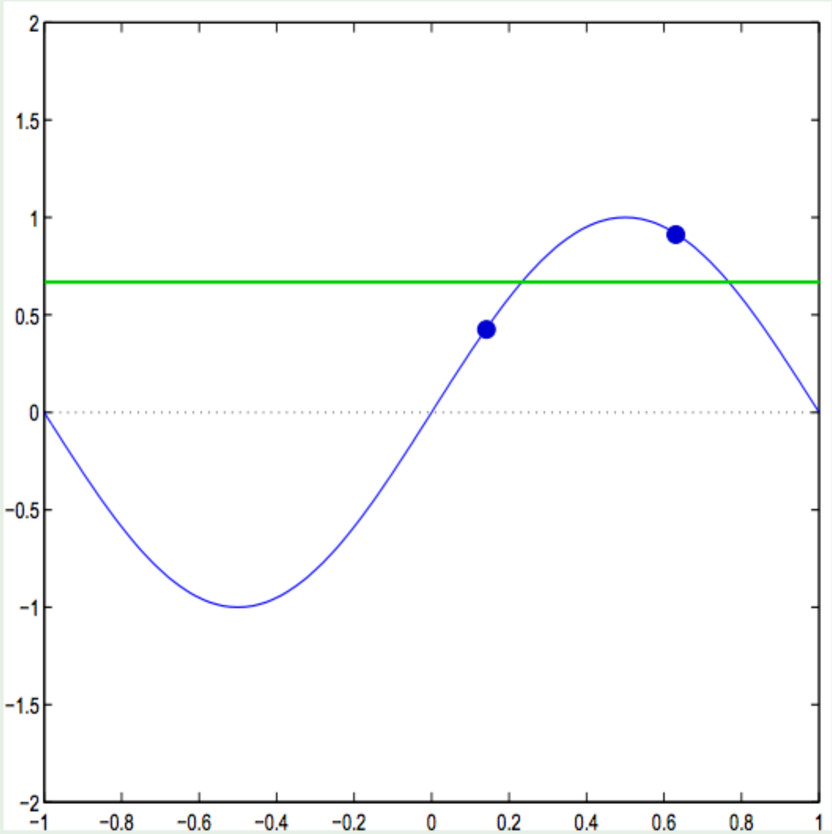


$\mathcal{H}_1$

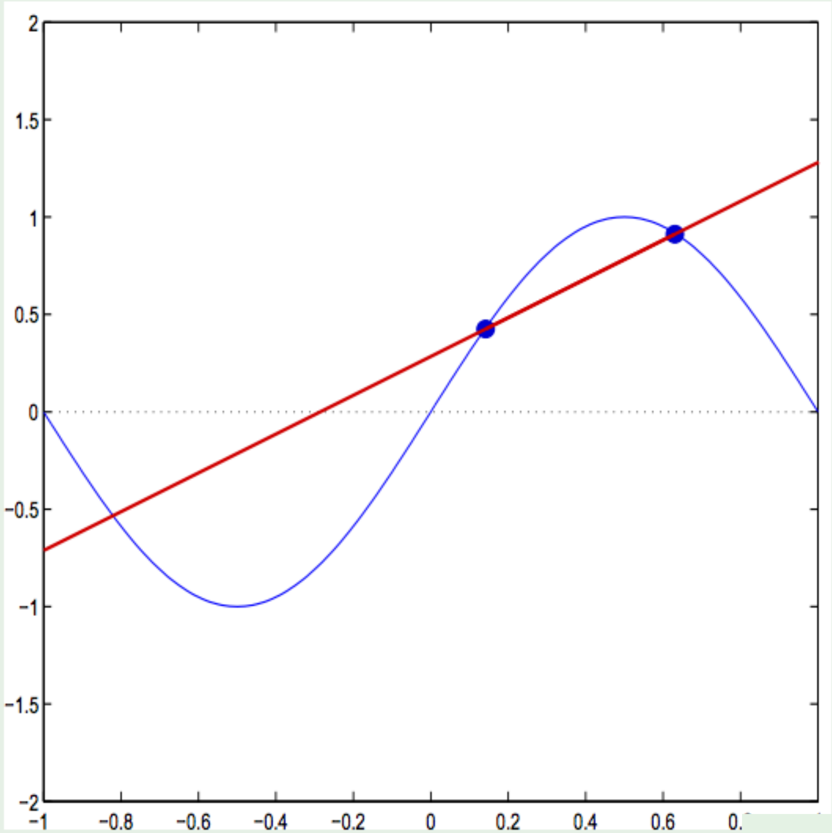


Learning - $\mathcal{H}_0$ versus $\mathcal{H}_1$

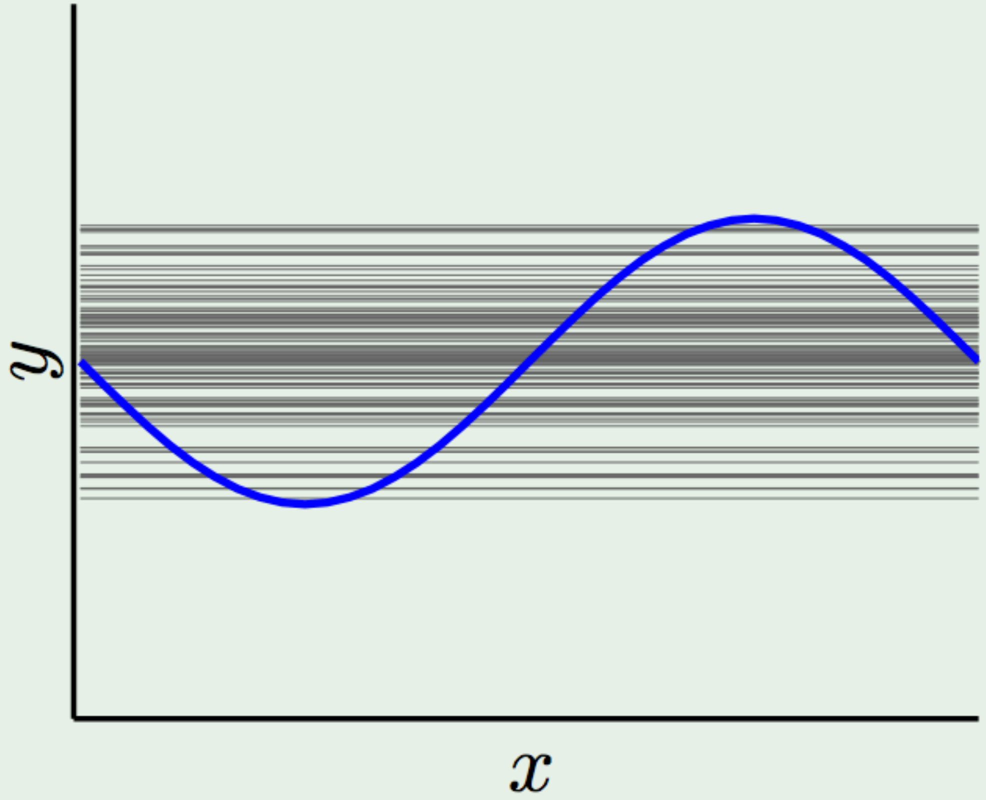
$\mathcal{H}_0$

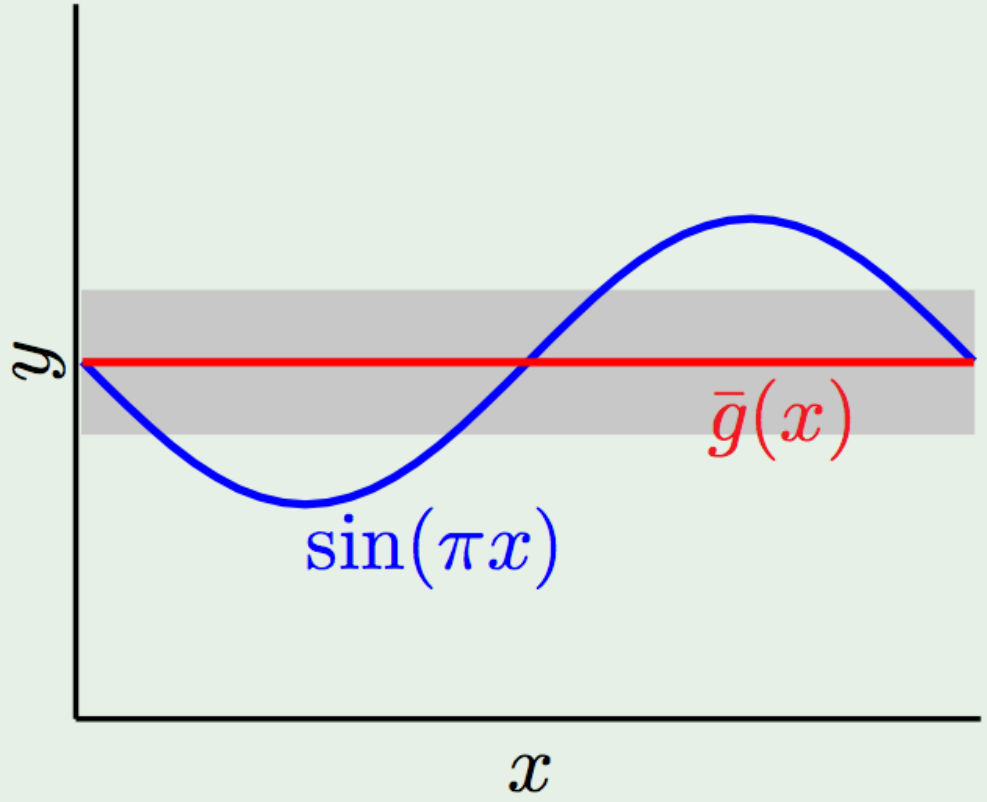


$\mathcal{H}_1$

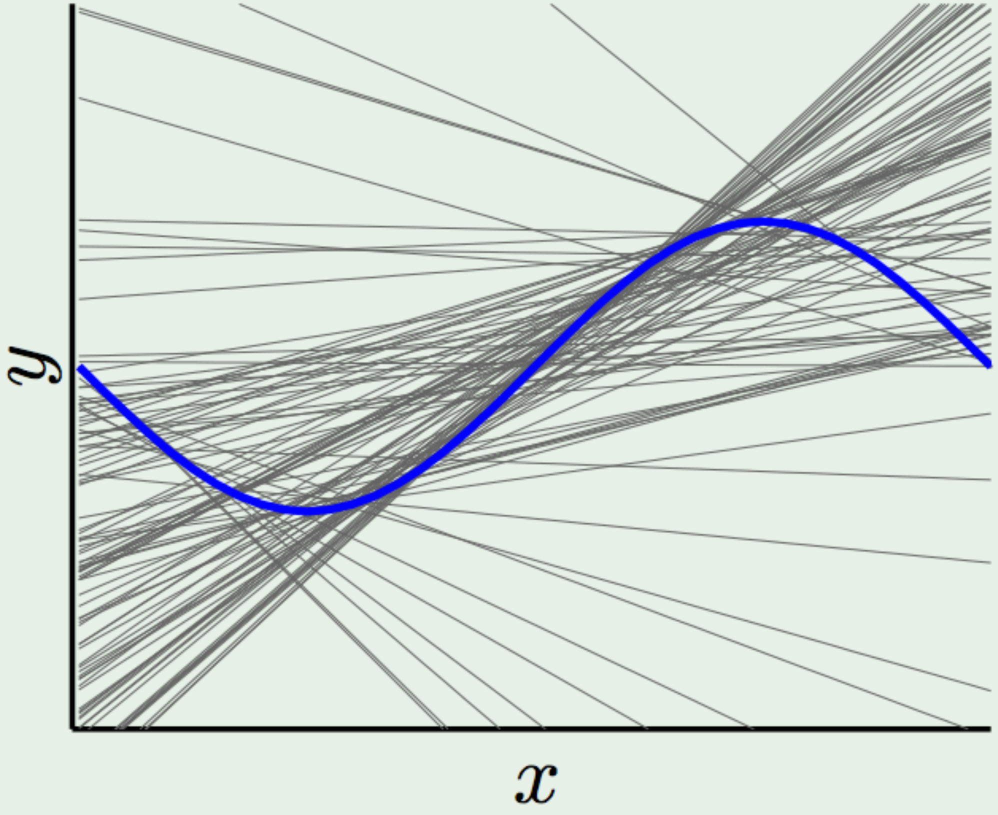


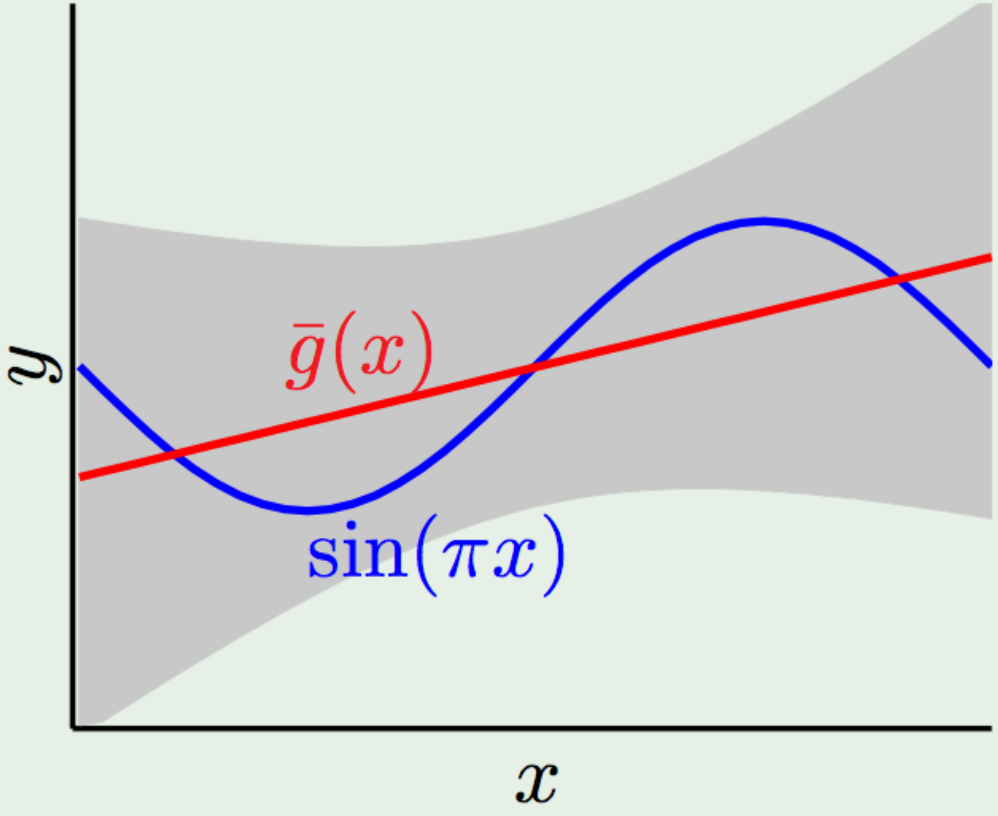
Bias and variance - $\mathcal{H}_0$





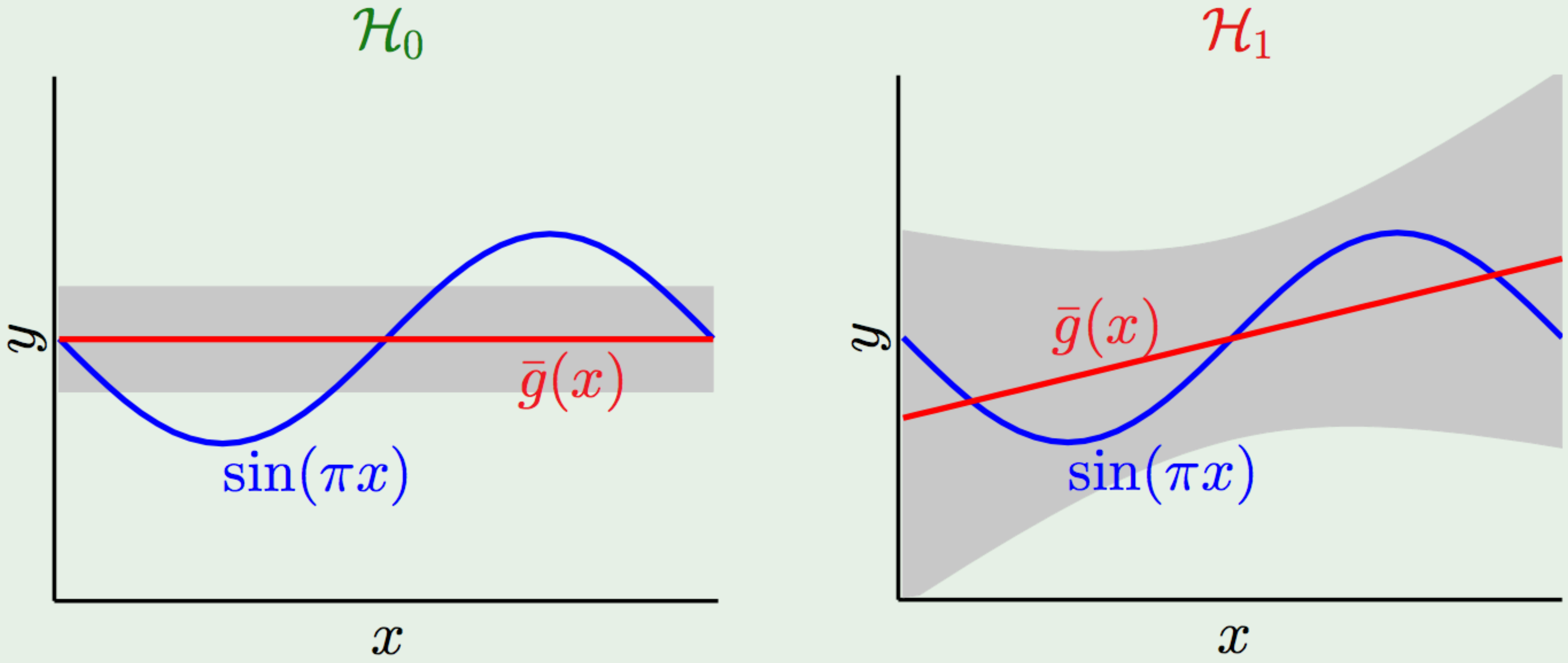
Bias and variance - $\mathcal{H}_1$





You will not be expected to compute these values

and the winner is . . .



bias $=$ **0.50** var $=$ **0.25**

bias $=$ **0.21** var $=$ **1.69**

Lesson learned

Match the **'model complexity'**

to the data resources, not to the target complexity

Outline

- Motivating example:
- Feature transformation
- Underfitting and overfitting
- Understanding error: Bias and varianc
- Learning curves
- validation and model selection
- Model selection (with limit
- K-fold cross validation

- Regularization

Yea!

Uh oh….

How to create a more complex hypothesis

Understanding what went wrong

Understanding where the error comes from, and how to estimate $E_{\text{out}}[g(\mathbf{x})]$

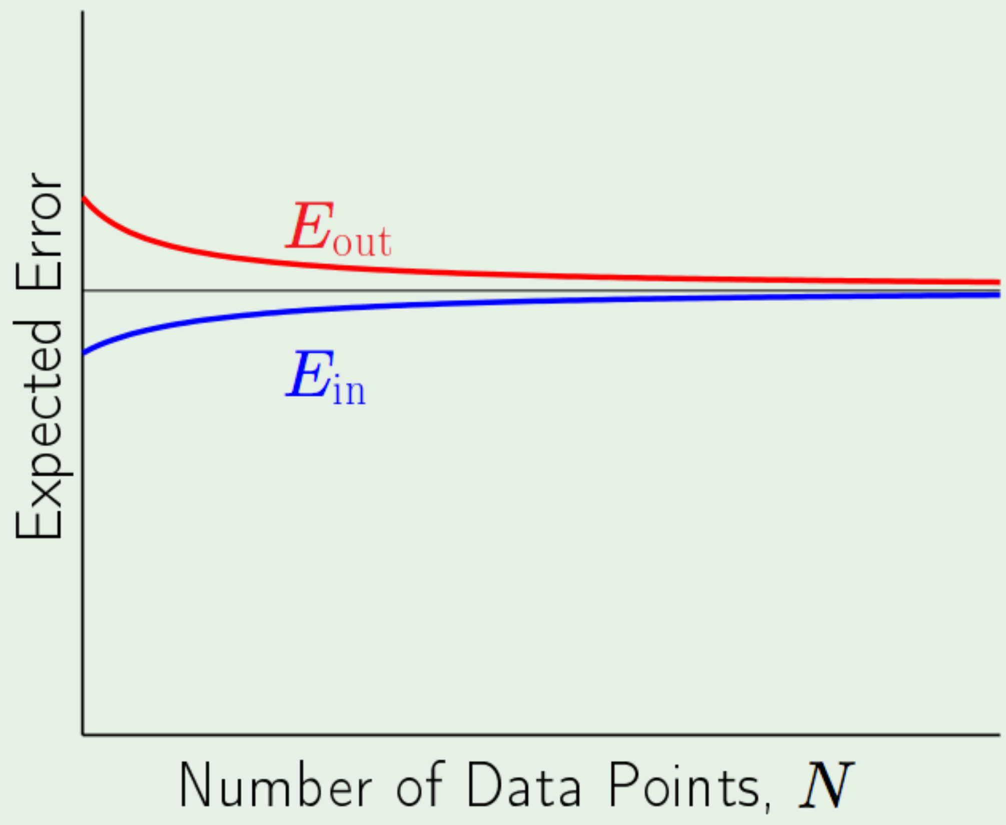
Our strategy

If we have many different hypothesis classes to choose from - how can we choose wisely?
And how can we estimate $E_{\text{out}}[g(\mathbf{x})]$?

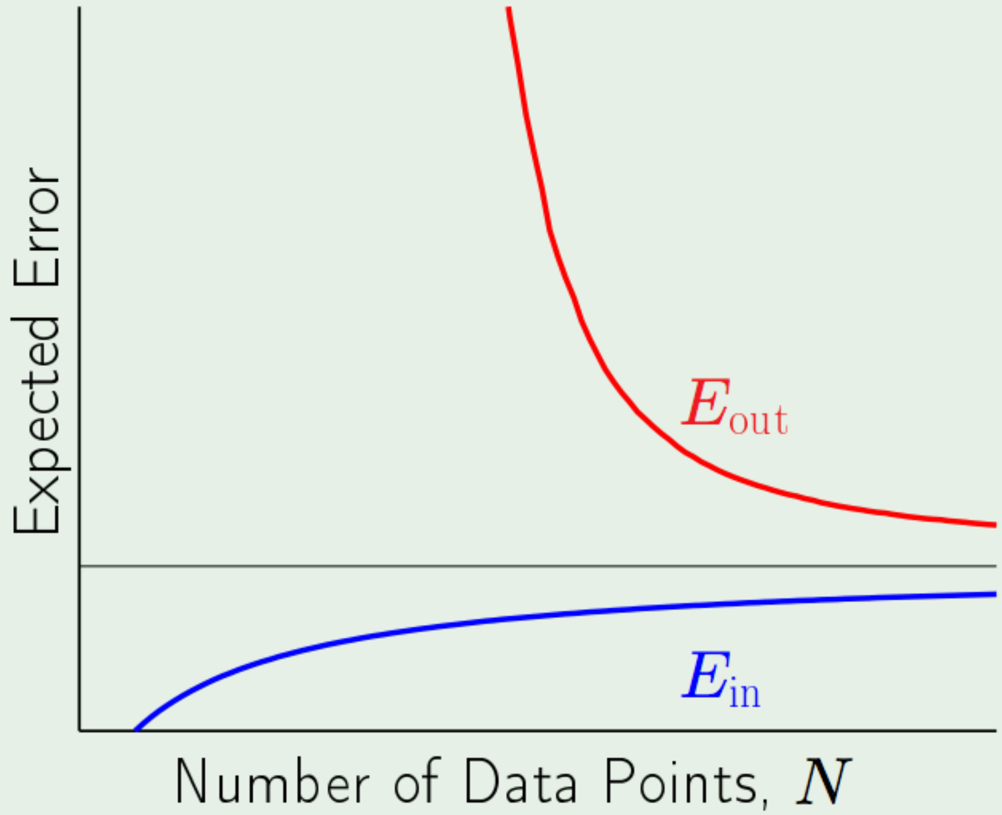
Pair Share

Do we expect the model to perform as well in the future as it performed on the training set?

The curves



Simple Model



Complex Model

Our goal is to minimize the generalization error (aka risk)

For linear regression, the goal is to minimize:

$$E_{\text{out}}(g(\mathbf{x})) = E[(y - g(\mathbf{x}))^2]$$

To do this we need to know the joint distribution of X and Y

How can we approximate this value?

Use our sample data!

…we could use our training examples to calculate our in-sample loss

$$E_{\text{in}}(g(\mathbf{x})) = \sum_{i=1}^{N} (y^{(i)} - g(\mathbf{x}^{(i)}))^2]$$

Empirical risk minimization by choosing the parameters with the highest likelihood

This is a very optimistic estimate!

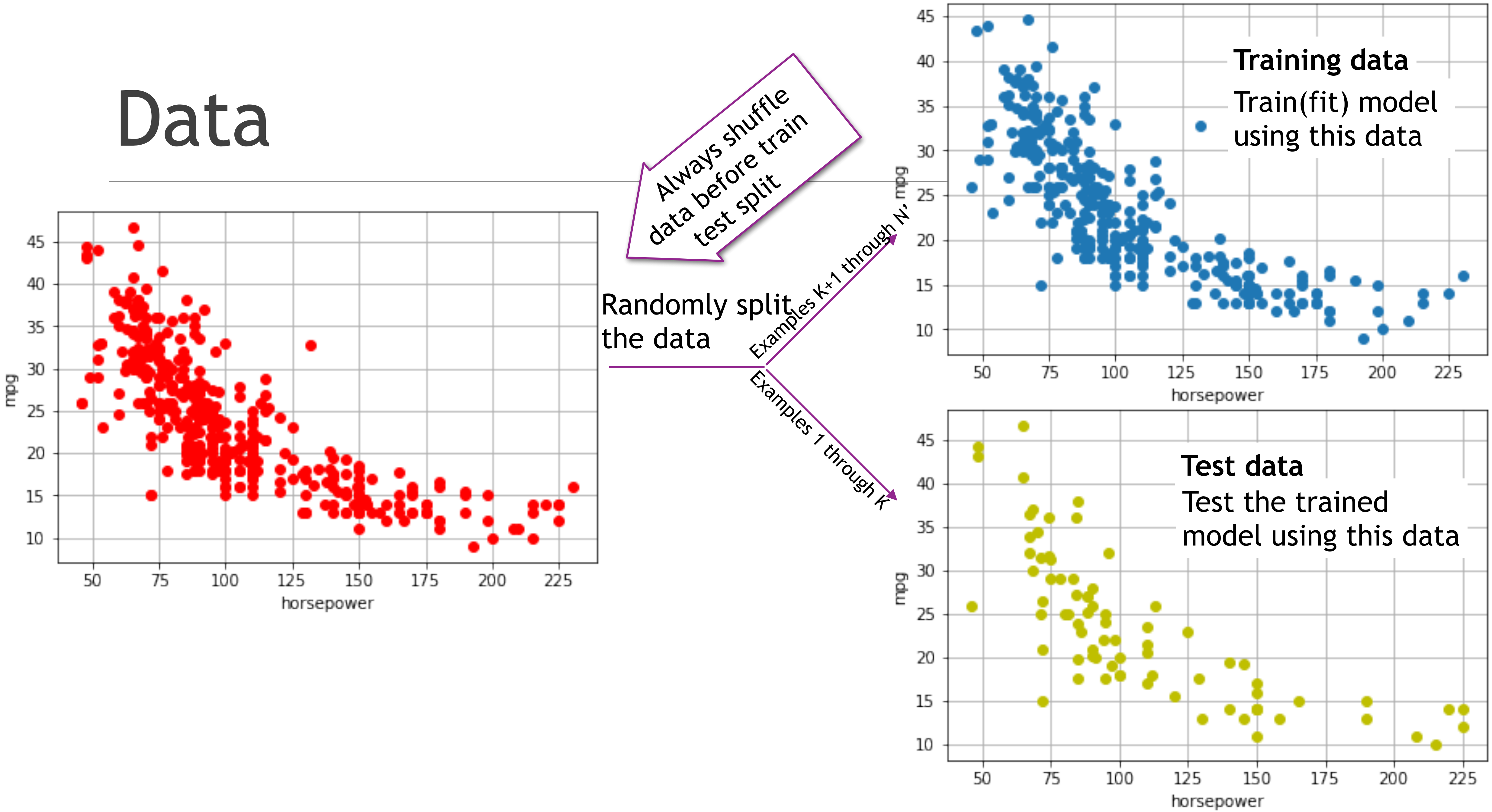
Pair share

The training error (cost) doesn't give the real world cost

$$E_{\text{out}}(g(\mathbf{x})) = E[(y - g(\mathbf{x}))^2]$$

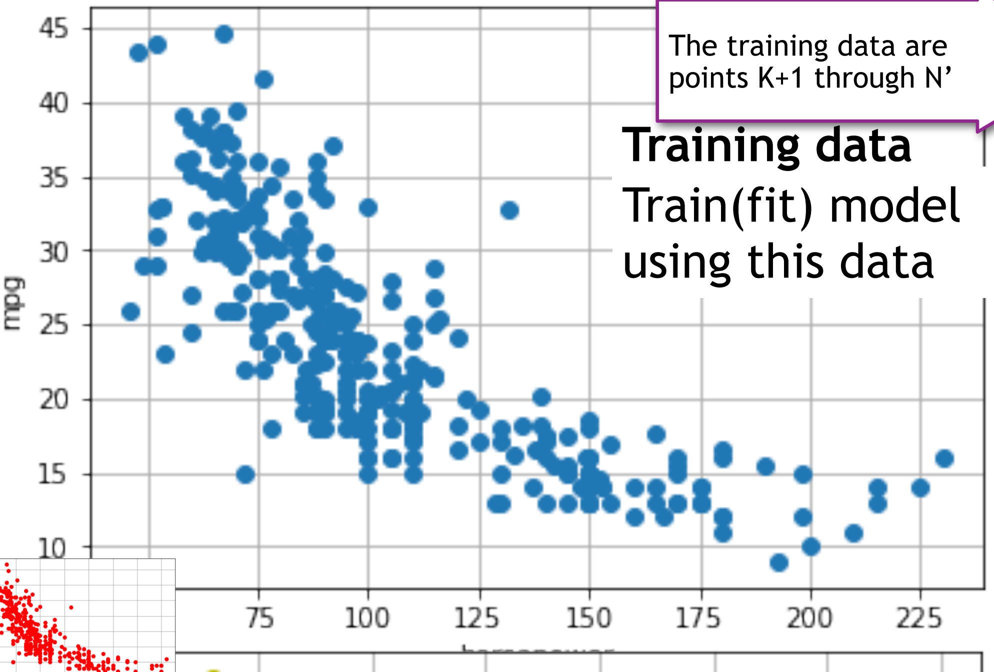
$$E_{\text{in}}(g(\mathbf{x})) << E_{\text{out}}(g(\mathbf{x}))$$

Data

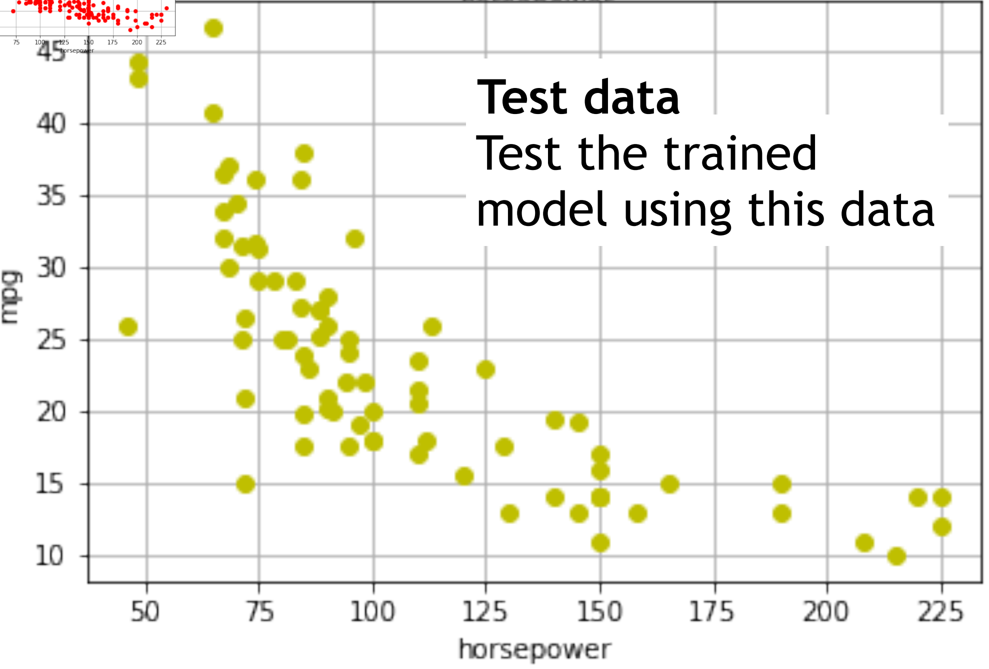


Always shuffle data before train test split
Randomly split the data
Examples K+1 through N'
Examples 1 through K
Training data
Train(fit) model using this data
Test data
Test the trained model using this data
mpg
horsepower

The training data are points K+1 through N’



Training data
Train(fit) model using this data



Test data
Test the trained model using this data

Fit model using the training data

Find the model that best fits **all** the training data

Determine $\hat{w}$ our estimated model parameters

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{|\text{training}|} \sum_{j \in \text{training}} \left(\text{mpg}^{(j)} - (w_0 + w_1 \text{horsepower}^{(j)}) \right)^2$$

Estimate the generalization error, $E_{\text{out}}(\text{w})$, by using the test data

$$E_{\text{test}}(\mathbf{w}) = \frac{1}{|\text{test}|} \sum_{j \in \text{test}} \left(\text{mpg}^{(j)} - (w_0 + w_1 \text{horsepower}^{(j)}) \right)^2$$

For binary classification, how good is our estimate for E_{out}

Is $|E_{out} - E_{test}|$ likely to be small?

“Hoeffding’s inequality is a powerful technique—perhaps the most important inequality in learning theory”

from http://cs229.stanford.edu/extra-notes/hoeffding.pdf

Generalization Bound for classification

iid: each example “has the same probability distribution as the others and all are mutually independent."

Suppose our test set contained K randomly chosen examples

then by using Hoeffding’s inequality

the probability our E_{out} differs from E_{test} by more than $\epsilon > 0$ occurs with probability at most $2e^{-2\epsilon^2 K}$

Example:

If K=500 and $\epsilon = 0.1$, then setting $\delta = 2e^{-2(0.1)^2(500)} = 0.0001$ then with probability $1 - \delta$ the true error is within 0.1 of the average error on the test set.

Generalization

Cannot get a range - instead get a confidence interval

Hoeffding inequality (stated without proof): for any sample size K, where each random variable is bounded in [a,b] the probability that the average value, v, of the random variables will deviate from its average μ by more than ϵ is:

$$P\left[\,|\,v - \mu\,| > \epsilon\right] \leq 2e^{-2\epsilon^2 K/(b-a)^2} = \delta \text{ for any } \epsilon > 0$$

Thus if $K \geq \dfrac{\log(2/\delta)(b-a)^2}{2\epsilon^2}$ then with probability $1-\delta$

v is ϵ close to μ

We are assuming the K examples are drawn iid from a distribution

Example:

Let g be a binary classifier (g outputs 0,1), let v be the average error of g on the test set of size K, and let μ be the true error of g. The probability that $|v - \mu| > \epsilon$ is at most $2e^{-2\epsilon^2 K}$

If K=500 and $\epsilon = 0.1$, then setting $\delta = 2e^{-2(0.1)^2(500)}$ then with probability $1 - \delta$ the true error is within 0.1 of the average error on the test set.

0.999909

Our estimated average error on our test set

Bound using numbers: K, ϵ and range of output values of function

Cannot get a range - instead get a confidence interval

Hoeffding inequality (stated without proof): for any sample size K, where each random variable is bounded in [a,b] the probability that the average value, v, of the random variables will deviate from its average μ by more than ϵ is:

$$P\left[\,|v-\mu|>\epsilon\right] \leq 2e^{-2\epsilon^2 K/(b-a)^2} = \delta \text{ for any } \epsilon > 0$$

Thus if $K \geq \frac{\log(2/\delta)(b-a)^2}{2\epsilon^2}$ then with probability $1-\delta$

v is ϵ close to μ

We are assuming the K examples are drawn iid from a distribution

Example:

Let g be a binary classifier (g outputs 0,1), let v be the average error of g on the test set of size K, and let μ be the true error of g. The probability that $|v-\mu| > \epsilon$ is at most $2e^{-2\epsilon^2 K}$

If K=500 and $\epsilon = 0.1$, then setting $\delta = 2e^{-2(0.1)^2(500)}$ then with probability $1-\delta$ the true error is within 0.1 of the average error on the test set.

0.999909

Generalization

Hoeffding inequality (stated without proof) for any sample size K, where each random variable is bounded in [a,b]] the probability that the average value, v, of the random variables will deviate from its average μ by more that ϵ is:

$$P\left[\,|v-\mu|>\epsilon\right]\leq 2e^{-2\epsilon^2K/(b-a)^2}=\delta \text{ for any } \epsilon>0$$

Thus if $K \geq \frac{\log(2/\delta)(b-a)^2}{2\epsilon^2}$ then with probability $1-\delta$

We are assuming the K examples are drawn iid from a distribution

v is ϵ close to μ

Example:

Let g be a binary classifier (g outputs 0,1), let v be the average error of g on the test set of size K, and let μ be the true error of g. The probability that $|v-\mu|>\epsilon$ is at most $2e^{-2\epsilon^2K}$

If K=100 and $\epsilon=0.2$, then $\underline{\delta=2e^{-2\cdot(0.2)^2\cdot 100}}$. With probability $1-\delta$ = 0.999 our estimated test set error is within 0.2 of the out of sample error