GENERAL ASSEMBLY

EVALUATING MODEL FIT

Sri Kanajan

LEARNING OBJECTIVES

- Define regularization, bias, and error metrics for regression problems
- Evaluate model fit using loss functions
- Select regression methods based on fit and complexity

COURSE

PRE-WORK

PRE-WORK REVIEW

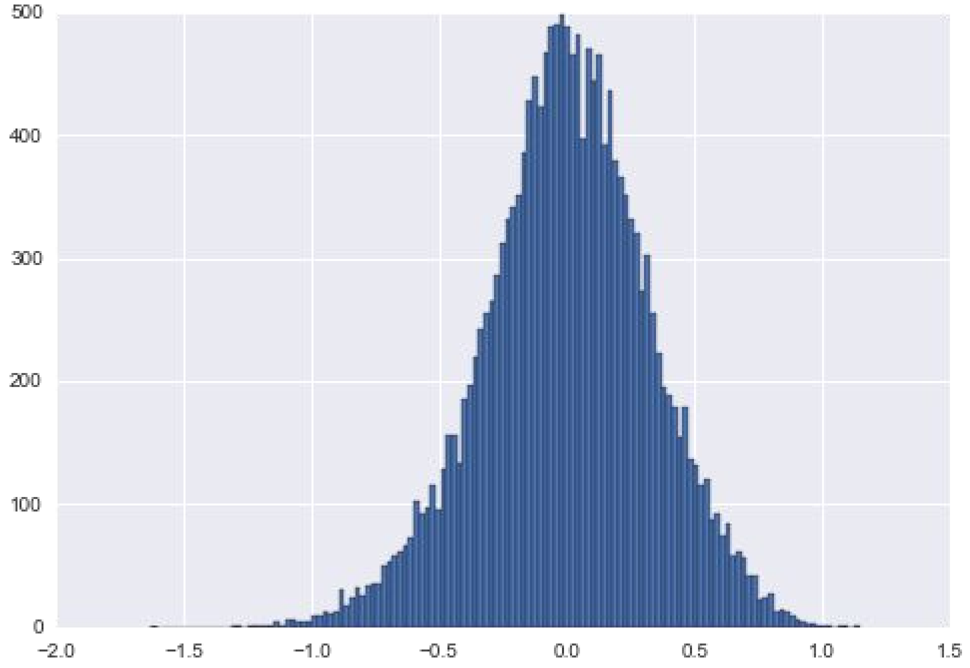
- Understand goodness of fit (r-squared)
- Measure statistical significance of feature coefficients
- Recall what a residual is
- Implement a sklearn estimator to predict a target variable

INTRODUCTION

LINEAR MODELS AND ERROR

RECALL: WHAT'S RESIDUAL ERROR?

- In linear models, residual error must be normal with a median close to zero.
- Individual residuals are useful to see the error of specific points, but it doesn't provide an overall picture for optimization.
- We need a metric to summarize the error in our model into one value.
- Mean square error: the mean residual error in our model



MEAN SQUARED ERROR (MSE)

- To calculate MSE:
    - Calculate the difference between each target y and the model's predicted value y-hat (i.e. the residual)
    - Square each residual.
    - Take the mean of the squared residual errors.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

MEAN SQUARED ERROR (MSE)

- sklearn's metrics module includes a mean_squared_error function.

```
from sklearn import metrics
metrics.mean_squared_error(y, model.predict(X))
```

MEAN SQUARED ERROR (MSE)

- For example, two arrays of the same values would have an MSE of 0.

```
from sklearn import metrics
metrics.mean_squared_error([1, 2, 3, 4, 5], [1, 2, 3, 4, 5])
0.0
```

MEAN SQUARED ERROR (MSE)

- Two arrays with different values would have a positive MSE.

```
from sklearn import metrics
metrics.mean_squared_error([1, 2, 3, 4, 5], [5, 4, 3, 2, 1])
# (4^2 + 2^2 + 0^2 + 2^2 + 4^2) / 5
8.0
```

HOW DO WE MINIMIZE ERROR?

- The regression method we've used is called "Ordinary Least Squares".
- This means that given a matrix X, solve for the *least* amount of square error for y.
- However, this assumes that X is unbiased, that it is representative of the population.

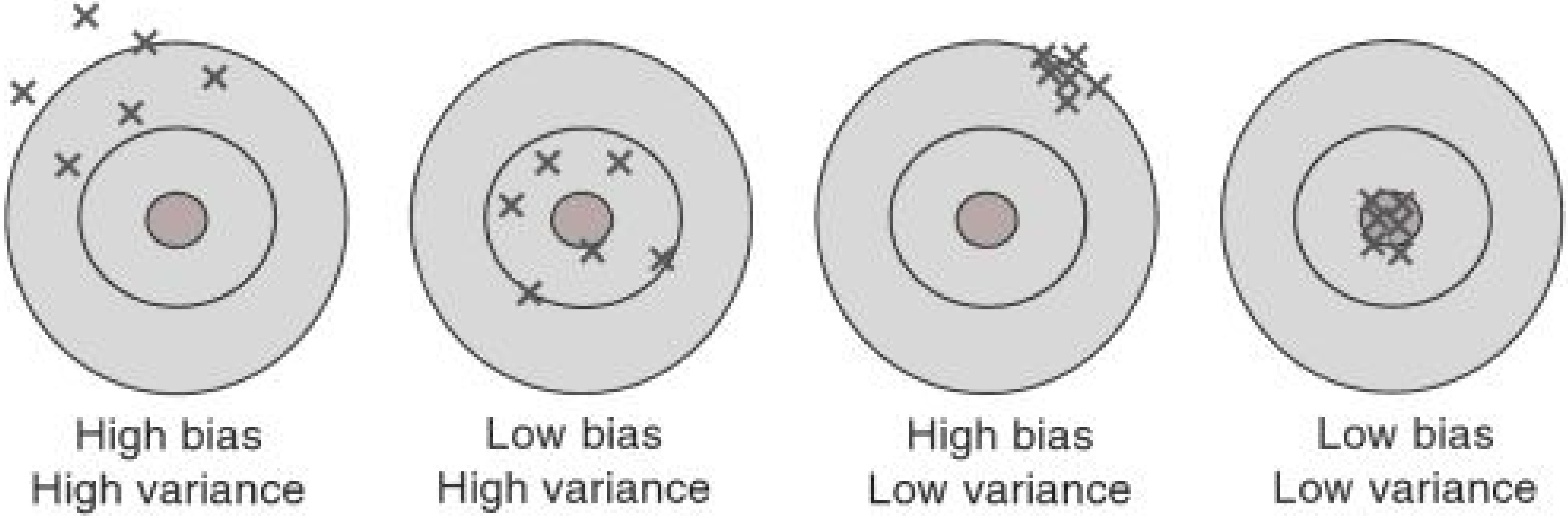
LET'S COMPARE TWO RANDOM MODELS

```python
import numpy as np
import pandas as pd
from sklearn import linear_model

df = pd.DataFrame({'x': range(100), 'y': range(100)})
biased_df  = df.copy()
biased_df.loc[:20, 'x'] = 1
biased_df.loc[:20, 'y'] = 1

def append_jitter(series):
    jitter = np.random.random_sample(size=100)
    return series + jitter
```

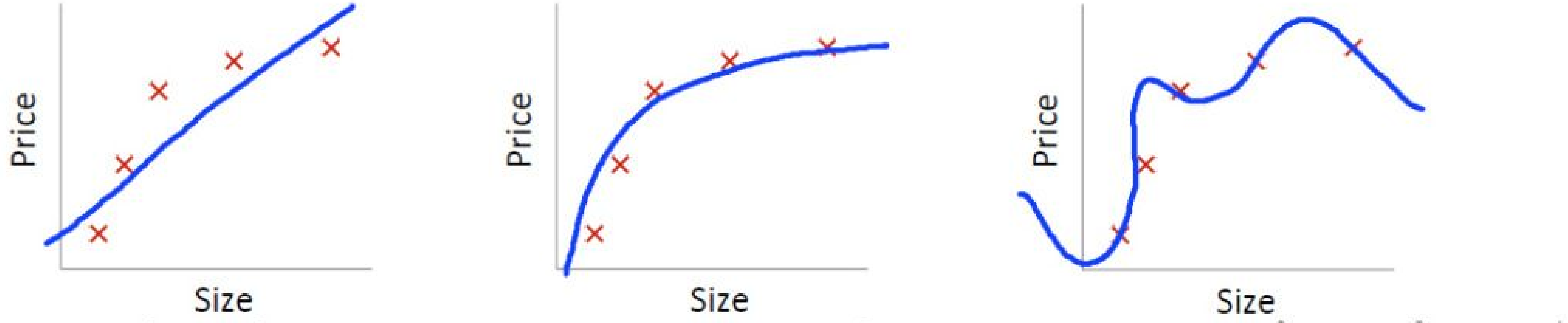
LET'S COMPARE TWO RANDOM MODELS

```
df['x'] = append_jitter(df.x)
df['y'] = append_jitter(df.y)

biased_df['x'] = append_jitter(biased_df.x)
biased_df['y'] = append_jitter(biased_df.y)

- Fit:
lm = linear_model.LinearRegression().fit(df[['x']], df['y'])
print metrics.mean_squared_error(df['y'], lm.predict(df[['x']]))

- Biased fit:
lm = linear_model.LinearRegression().fit(biased_df[['x']], biased_df['y'])
print metrics.mean_squared_error(df['y'], lm.predict(df[['x']]))
```

BIAS VS. VARIANCE



BIAS VS. VARIANCE



$\theta_0 + \theta_1 x$

High bias
(underfit)

$d=1$

$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

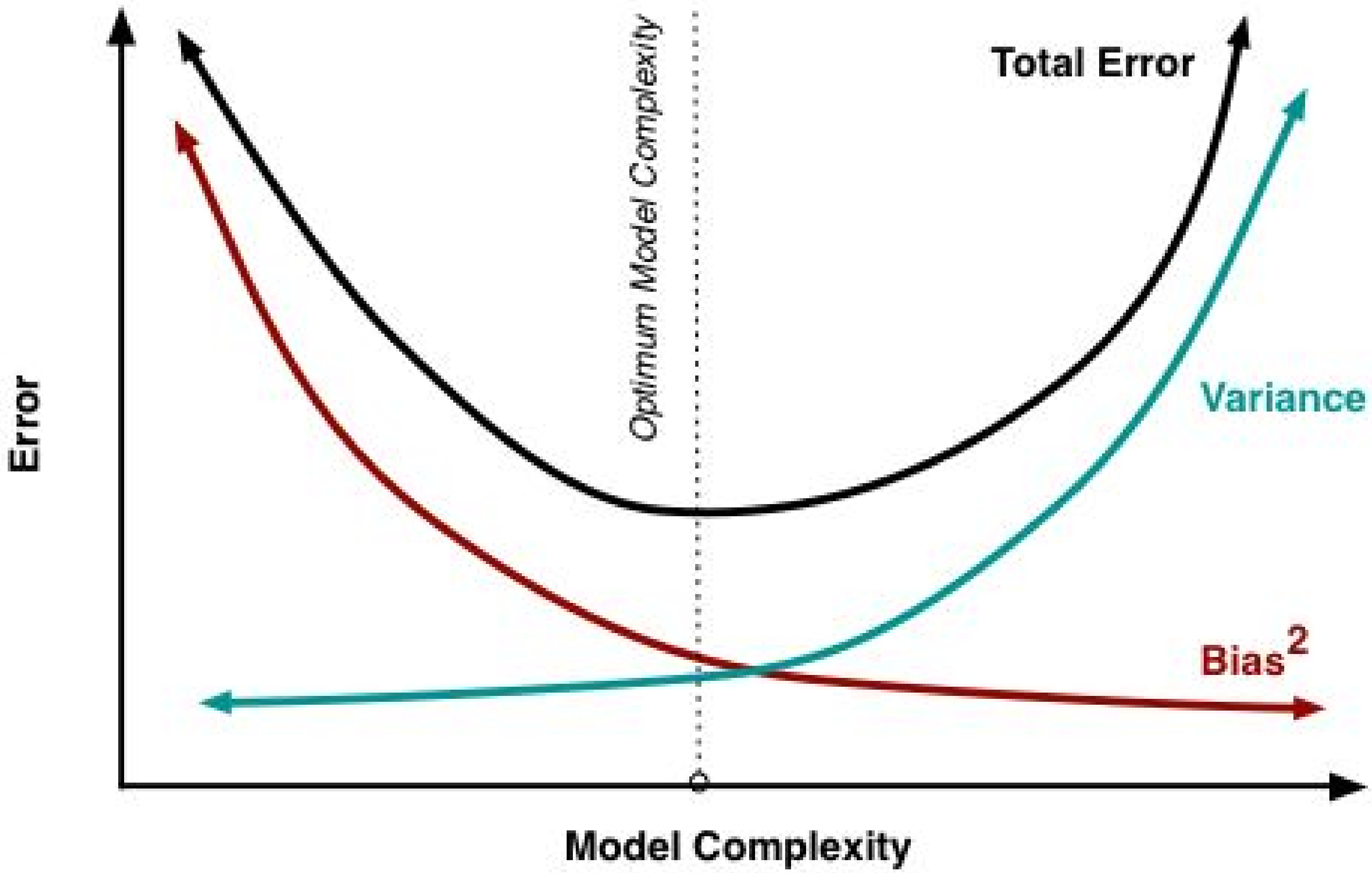
$d=2$

$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance
(overfit)

$d=4$

BIAS VARIANCE TRADEOFF



DEMO

CROSS VALIDATION

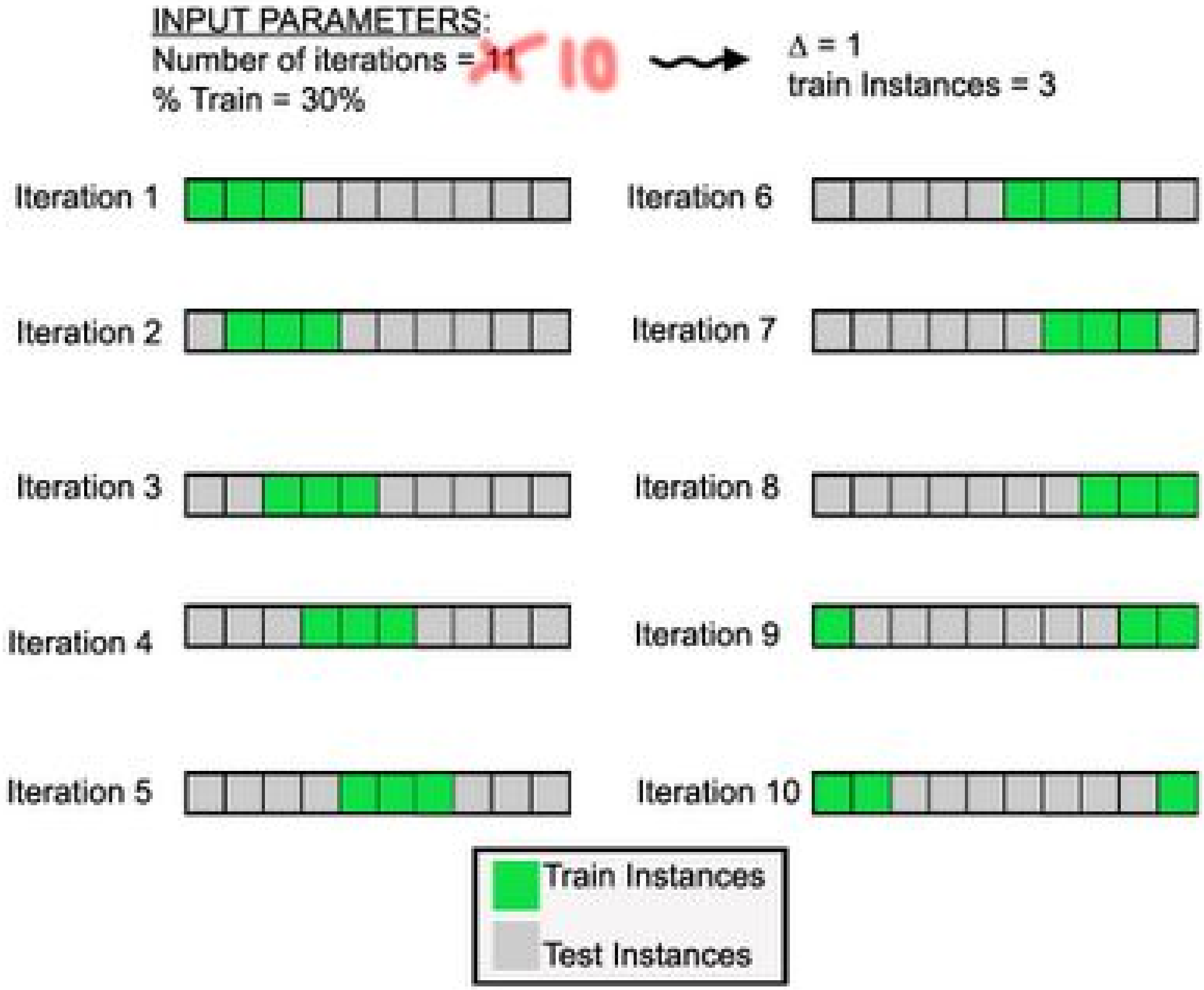
CROSS VALIDATION

- Cross validation can help account for bias.
- The general idea is to
  - Generate several models on different cross sections of the data
  - Measure the performance of each
  - Take the mean performance
- This technique swaps bias error for generalized error, describing previous trends accurately enough to extend to future trends.

K-FOLD CROSS VALIDATION

- k-fold cross validation
  - Split the data into *k* group
  - Train the model on all segments except one
  - Test model performance on the remaining set
- If k = 5, split the data into five segments and generate five models.

CROSS VALIDATION



USING K-FOLD CROSS VALIDATION WITH MSE

- Import the appropriate packages and load data.

```
from sklearn import cross_validation
wd = '../../datasets/'
bikeshare = pd.read_csv(wd + 'bikeshare/bikeshare.csv')
weather = pd.get_dummies(bikeshare.weathersit, prefix='weather')
modeldata = bikeshare[['temp', 'hum']].join(weather[['weather_1',
'weather_2', 'weather_3']])
y = bikeshare.casual
```

USING K-FOLD CROSS VALIDATION WITH MSE

- Build models on subsets of the data and calculate the average score.

```
kf = cross_validation.KFold(len(modeldata), n_folds=5, shuffle=True)
scores = []
for train_index, test_index in kf:
    lm = linear_model.LinearRegression().fit(modeldata.iloc
[train_index], y.iloc[train_index])
    scores.append(metrics.mean_squared_error(y.iloc[test_index], lm.
predict(modeldata.iloc[test_index])))

print np.mean(scores)
```

USING K-FOLD CROSS VALIDATION WITH MSE

- This can be compared to the model built on all of the data.

```
- This score will be lower, but we're trading off bias error for
generalized error:
lm = linear_model.LinearRegression().fit(modeldata, y)
print metrics.mean_squared_error(y, lm.predict(modeldata))
```

- Which approach would predict new data more accurately?

GUIDED PRACTICE

CROSS VALIDATION WITH LINEAR REGRESSION

ACTIVITY: CROSS VALIDATION WITH LINEAR REGRESSION

EXERCISE

DIRECTIONS (20 minutes)

If we were to continue increasing the number of folds in cross validation, would error increase or decrease?

1. Using the previous code example, perform k-fold cross validation for all even numbers between 2 and 50.
2. Answer the following questions:
    a. What does `shuffle=True` do?
    b. At what point does cross validation no longer seem to help the model?
3. Hint: `range(2, 51, 2)` produces a list of even numbers from 2 to 50

DELIVERABLE

Answers to questions

INTRODUCTION

REGULARIZATION AND CROSS VALIDATION

WHAT IS REGULARIZATION? AND WHY DO WE USE IT?

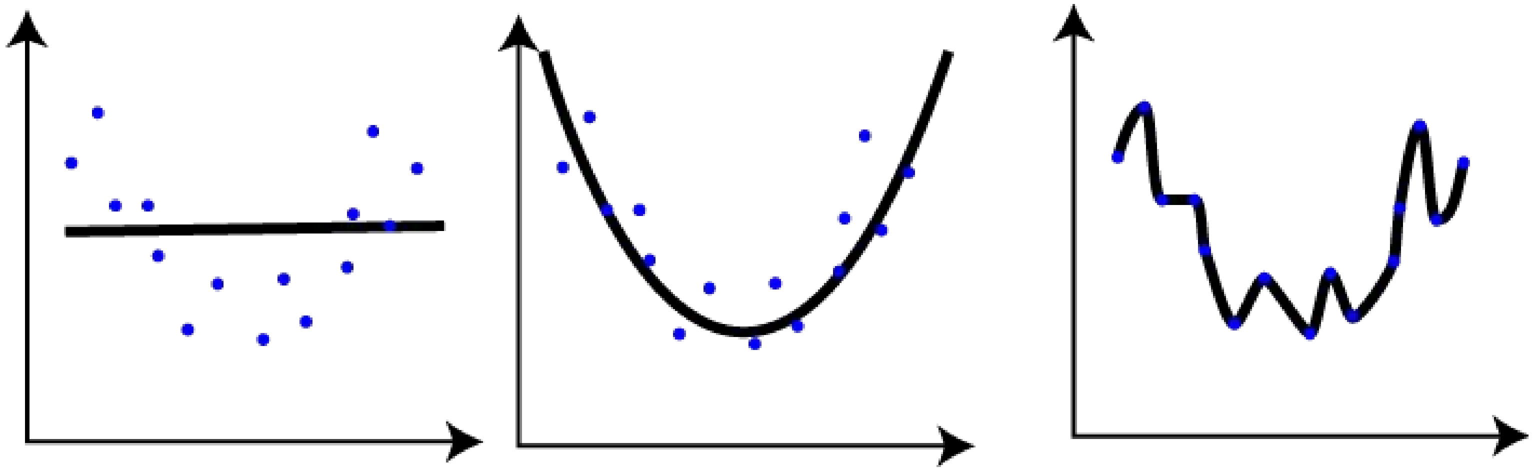
- Regularization is an additive approach to protect models against overfitting (being potentially biased and overconfident, not generalizing well).

- Regularization becomes an additional weight to coefficients, shrinking them closer to zero.

- L1 (Lasso Regression) adds the extra weight to the sum of the absolute of the coefficients.

- L2 (Ridge Regression) adds the weight to the sum of the square of the coefficients.
-
- Use Lasso when you want less features (sparse model) and Ridge when you want the most R^2 but are okay with more features. Lasso will also drop features when they are somewhat correlated to each other

WHAT IS OVERFITTING?



- The first model poorly explains the data.
- The second model explains the general curve of the data.
- The third model drastically overfits the model, bending to every point.
- Regularization helps prevent the third model.

WHERE REGULARIZATION MAKES SENSE

- What happens to MSE if use Lasso or Ridge Regression directly?

```
lm = linear_model.LinearRegression().fit(modeldata, y)
print metrics.mean_squared_error(y, lm.predict(modeldata))
lm = linear_model.Lasso().fit(modeldata, y)
print metrics.mean_squared_error(y, lm.predict(modeldata))
lm = linear_model.Ridge().fit(modeldata, y)
print metrics.mean_squared_error(y, lm.predict(modeldata))
```

```
1672.58110765 # OLS
1725.41581608 # L1
1672.60490113 # L2
```

WHERE REGULARIZATION MAKES SENSE

- It doesn’t seem to help. Why is that?
- We need to optimize the regularization weight parameter (called alpha) through cross validation.

ACTIVITY: KNOWLEDGE CHECK

EXERCISE

ANSWER THE FOLLOWING QUESTIONS (5 minutes)

1. Why is regularization important?
2. What does it protect against and how?

DELIVERABLE

Answers to the above questions

DEMO

UNDERSTANDING REGULARIZATION EFFECTS

QUICK CHECK

- We are working with the bikeshare data to predict riders over hours/days with a few features.

- Does it make sense to use a ridge regression or a lasso regression?

- Why?

UNDERSTANDING REGULARIZATION EFFECTS

- Let's test a variety of alpha weights for Ridge Regression on the bikeshare data.

```
alphas = np.logspace(-10, 10, 21)
for a in alphas:
    print 'Alpha:', a
    lm = linear_model.Ridge(alpha=a)
    lm.fit(modeldata, y)
    print lm.coef_
    print metrics.mean_squared_error(y, lm.predict(modeldata))
```

- What happens to the weights of the coefficients as alpha increases? What happens to the error as alpha increases?

WE CAN MAKE THIS EASIER WITH GRID SEARCH!

- Grid search exhaustively searches through all given options to find the best solution. Grid search will try all combos given in `param_grid`.

```
param_ grid = {
    'intercept': [True, False],
    'alpha': [1, 2, 3],
}
```

WE CAN MAKE THIS EASIER WITH GRID SEARCH!

- This param grid has six different options:
  - intercept True, alpha 1
  - intercept True, alpha 2
  - intercept True, alpha 3
  - intercept False, alpha 1
  - intercept False, alpha 2
  - intercept False, alpha 3

```
param_ grid = {
    'intercept': [True, False],
    'alpha': [1, 2, 3],
}
```

WE CAN MAKE THIS EASIER WITH GRID SEARCH!

- This is an incredibly powerful, automated machine learning tool!

```
from sklearn import grid_search

alphas = np.logspace(-10, 10, 21)
gs = grid_search.GridSearchCV(
    estimator=linear_model.Ridge(),
    param_grid={'alpha': alphas},
    scoring='mean_squared_error')
```

WE CAN MAKE THIS EASIER WITH GRID SEARCH!

```
gs.fit(modeldata, y)

print -gs.best_score_ # mean squared error here comes in negative, so let's make it positive.
print gs.best_estimator_ # explains which grid_search setup worked best
print gs.grid_scores_ # shows all the grid pairings and their performances.
```

GUIDED PRACTICE

GRID SEARCH CV, SOLVING FOR ALPHA

ACTIVITY: GRID SEARCH CV, SOLVING FOR ALPHA

EXERCISE

DIRECTIONS (25 minutes)

1. Modify the previous code to do the following:
    a. Introduce cross validation into the grid search. This is accessible from the cv argument.
    b. Add fit_intercept = True and False to the param_grid dictionary.
    c. Re-investigate the best score, best estimator, and grid score attributes as a result of the grid search.

DELIVERABLE

New code and output that meets above requirements

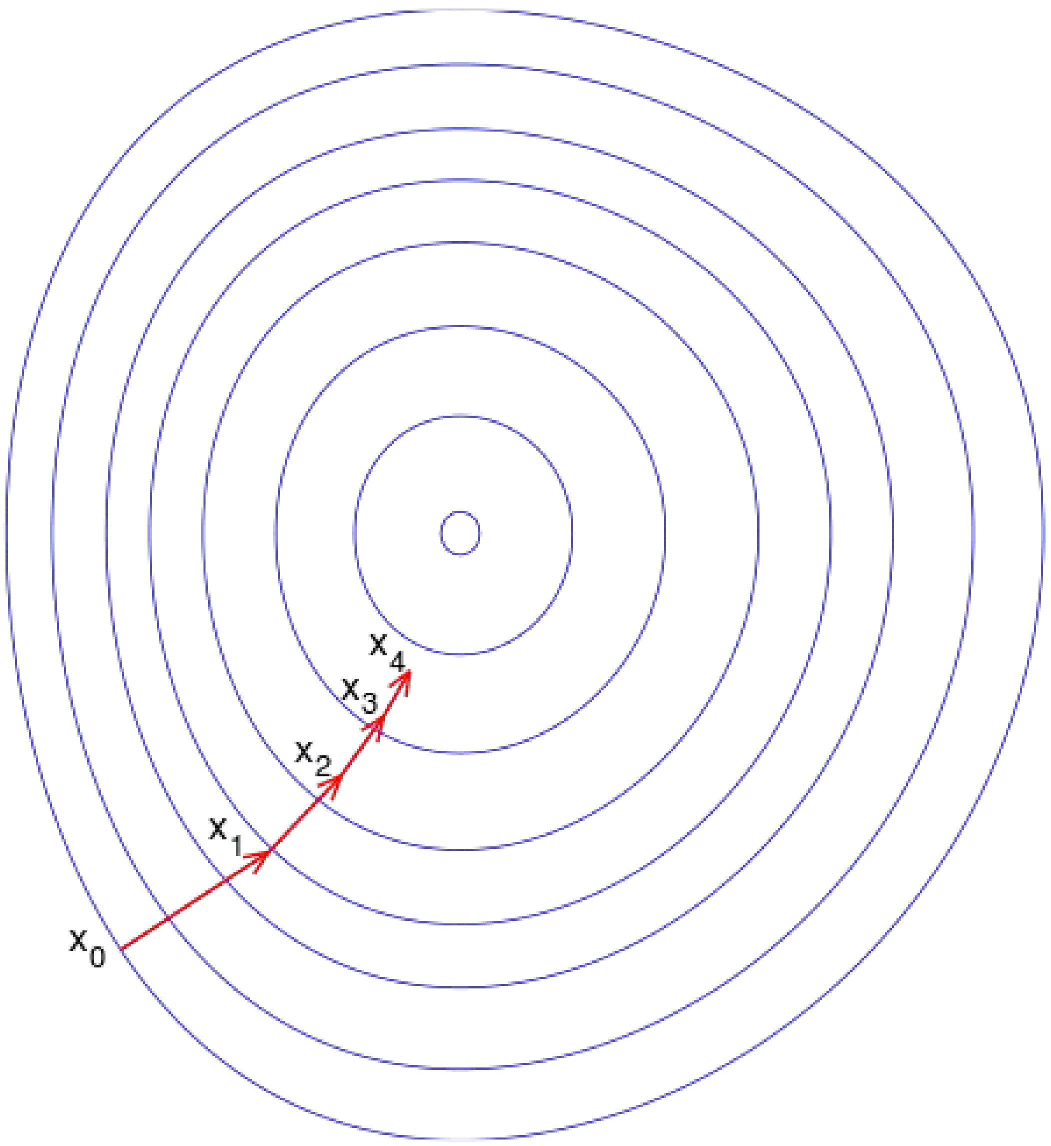
MINIMIZING LOSS THROUGH GRADIENT DESCENT

GRADIENT DESCENT

- Gradient Descent can also help us minimize error.

- How Gradient Descent works:
  - A random linear solution is provided as a starting point
  - The solver attempts to find a next “step”: take a step in any direction and measure the performance.
  - If the solver finds a better solution (i.e. lower MSE), this is the new starting point.
  - Repeat these steps until the performance is optimized and no “next steps” perform better. The size of steps will shrink over time.

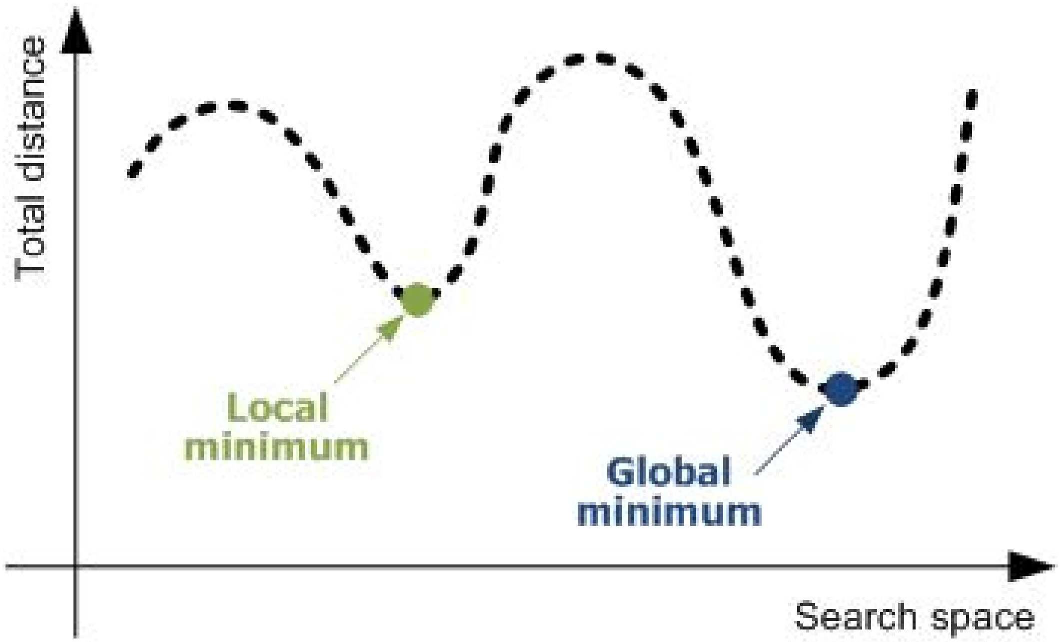
GRADIENT DESCENT



GLOBAL VS LOCAL MINIMUMS

- Gradient Descent could solve for a *local* minimum instead of a *global* minimum.

- A *local* minimum is confined to a very specific subset of solutions. The *global* minimum considers all solutions. These could be equal, but that's not always true.



INDEPENDENT PRACTICE

ON YOUR OWN

ACTIVITY: ON YOUR OWN

EXERCISE

DIRECTIONS (30 minutes)

There are tons of ways to approach a regression problem.

1. Demonstrate the grid_search module.
2. Use a model you evaluated last class or the simpler one from today. Implement param_grid in grid search to answer the following questions:
   a. With a set of values between 10^-10 and 10^-1, how does MSE change?
   b. Our data suggests we use L1 regularization. Using a grid search with l1_ratios between 0 and 1, increasing every 0.05, does this statement hold true?

DELIVERABLE

Answered questions

ACTIVITY: ON YOUR OWN

EXERCISE

Starter Code

```
params = {} # put your gradient descent parameters here
gs = grid_search.GridSearchCV(
    estimator=linear_model.SGDRegressor(),
    cv=cross_validation.KFold(len(modeldata), n_folds=5, shuffle=True),
    param_grid=params,
    scoring='mean_squared_error',
    )

gs.fit(modeldata, y)

print 'BEST ESTIMATOR'
print -gs.best_score_
print gs.best_estimator_
print 'ALL ESTIMATORS'
print gs.grid_scores_
```

CONCLUSION

TOPIC REVIEW

LESSON REVIEW

- What's the (typical) range of r-squared?
- What's the range of mean squared error?
- How would changing the scale or interpretation of y (your target variable) effect mean squared error?
- What's cross validation, and why do we use it in machine learning?
- What is error due to bias? What is error due to variance? Which is better for a model to have, if it had to have one?
- How does gradient descent try a different approach to minimizing error?

COURSE

Project Presentations

BEFORE NEXT CLASS

- Project: Final Project, Deliverable 1

LESSON

Q & A

LESSON

EXIT TICKET

DON'T FORGET TO FILL OUT YOUR EXIT TICKET