LINEAR REGRESSION

March 05, 2015 – General Assembly, Santa Monica

Mohsen Chitsaz, Ph.D.
Sr. Principal Data Scientist at Symantec

Lecture 08

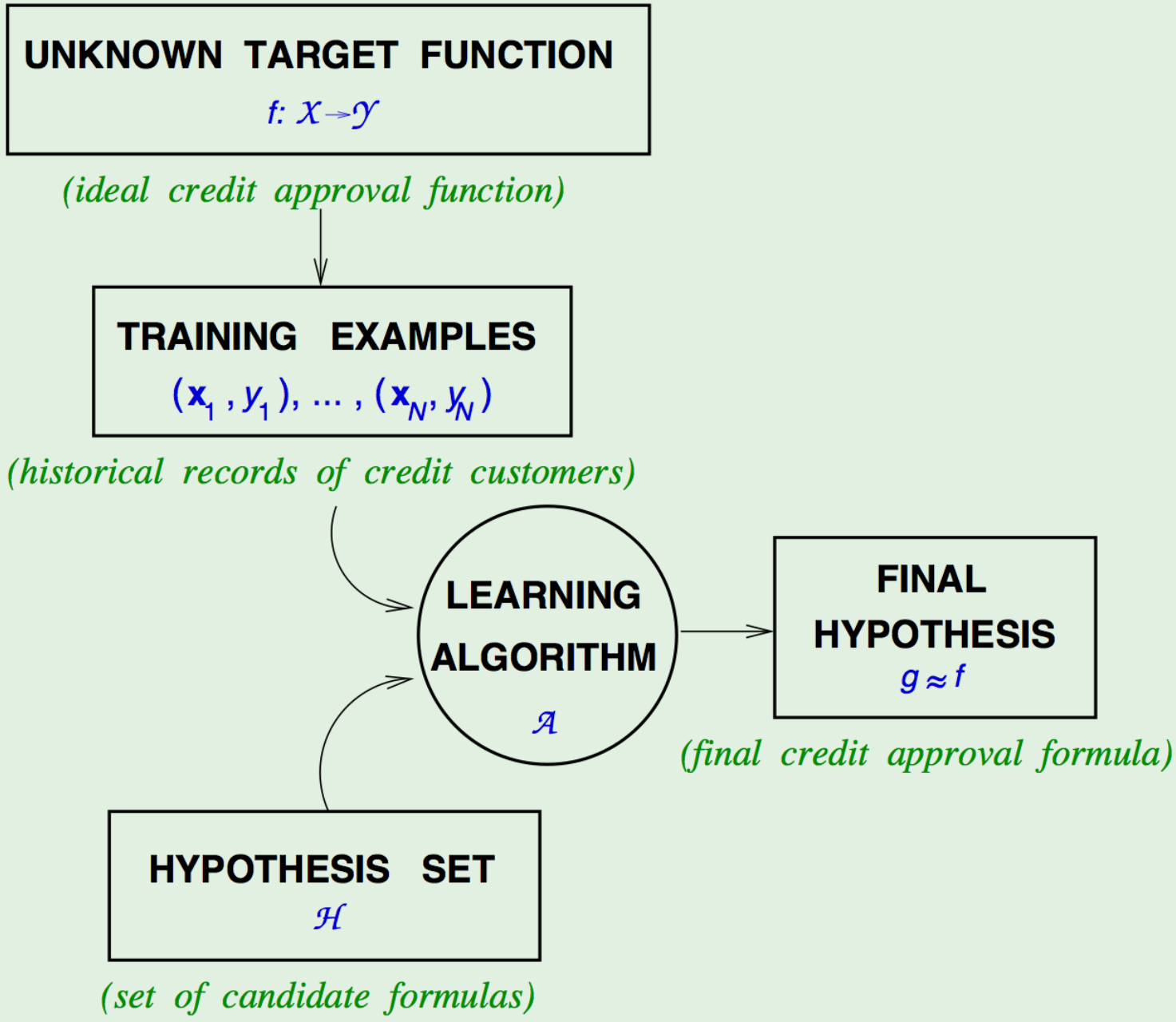
Machine Learning

- **Supervised**
  - Regression
  - Classification
- **Unsupervised**
  - Clustering
  - Density estimation
  - Dimensionality reduction

Components of Learning

- Input: $\mathbf{x}$ (*customer application*)
- Output: y (*good/bad customer?*)
- Target function: $f : \mathcal{X} \rightarrow \mathcal{Y}$ (*ideal credit approval formula*)
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$ (*historical records*)

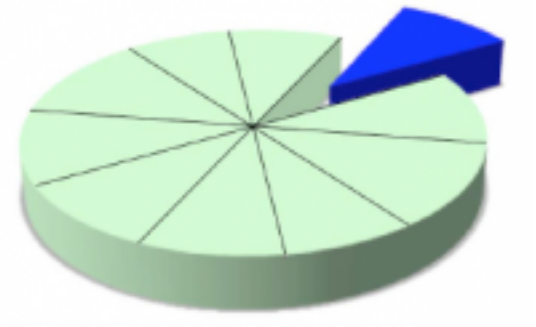
- Hypothesis: $g : \mathcal{X} \rightarrow \mathcal{Y}$ (*formula to be used*)


UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \to \mathcal{Y}$
(ideal credit approval function)
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
(historical records of credit customers)
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g \approx f$
(final credit approval formula)
HYPOTHESIS SET
$\mathcal{H}$
(set of candidate formulas)

Cross Validation (Rotation Estimation)

Cross Validation

- K-fold cross validation
  - Splitting the data into K pieces
  - Repeating the process of training K-times
    - For each training round, use (K-1) pieces and leave one piece out
    - Train on (K-1) parts and test on the piece that is left out



K-Fold Cross Validation

```
folds = cv.KFold(n=114, n_folds=6)
for fold in folds:
  train_data = xl.ix[fold[0],:]
  test_data = xl.ix[fold[1],:]
```

Regression

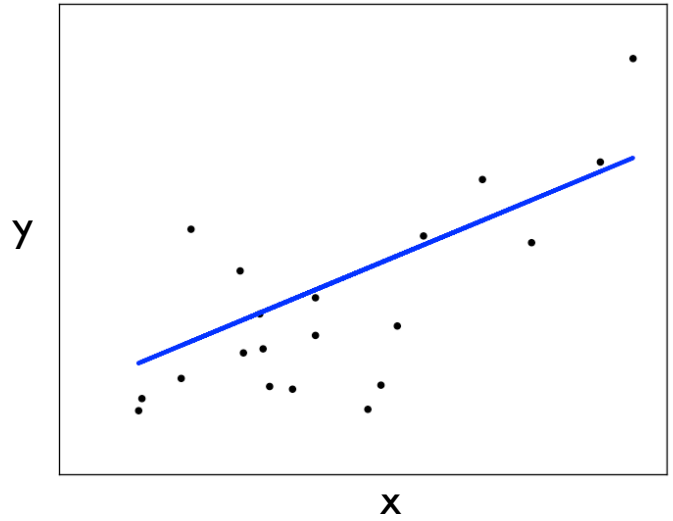
- Linear Regression

- Polynomial Regression

Linear Regression

Generalized Linear Models

$$\hat{y}(w, x) = \underbrace{w_0}_{\text{Intercept}} + \underbrace{w_1 x_1 + \ldots + w_p x_p}_{\text{weighted inputs}}$$



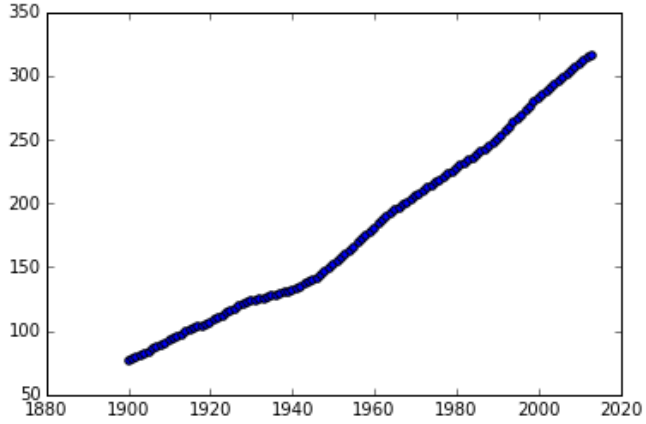
Ordinary Least Squares

- Minimizes the error,
  X: input vector, **y:** output value, **w:** weights

$$\min_{w} ||Xw - y||_2^2$$

Regression Example – US Population

Year	Population
1900	76.09
1901	77.58
1902	79.16
1903	80.63
1904	82.17
1905	83.82
1906	85.45
1907	87.01
1908	88.71



How good is our model?

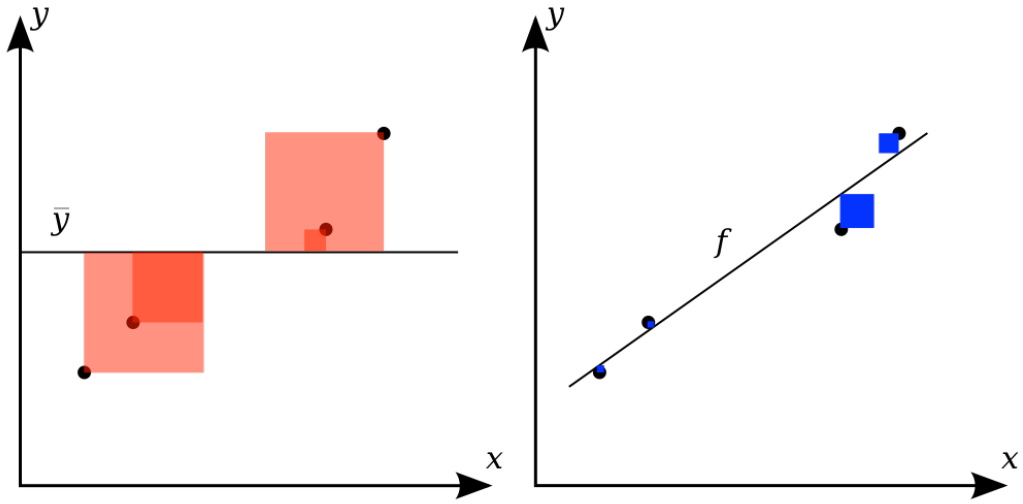
- We need to know how well the model fits the data

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2$$

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$



`sklearn.metrics.r2_score`

Polynomial Regression

- The goal of polynomial regression is to model a non-linear relationship between the independent and dependent variables

- Example for one dimensional

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n$$

Polynomial Regression

- Still linear regression

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n$$

- Sklearn:

```
import sklearn.preprocessing as pp
poly = pp.PolynomialFeatures()
b = np.arange(0,6).reshape(2,3)
c = poly.fit_transform(b)
```

Class Group Work – Abalone Age

- Determining the age Abalone is very laborious
- We want to find a formula that predicts the **age** of abalone based on some of its features

Class Group Work – Abalone Age

Feature	Description
sex	M, F, I, (Gender or Infant)
length	Longest shell measurement (mm)
diameter	Perpendicular to the length (mm)
height	With meat in shell (mm)
whole_weight (gr)	Whole weight (gr)
shucked_weight	Weight of meat (gr)
viscera_weight	Gut weight after bleeding (gr)
shell_weight	After being dried (gr)
rings	+1.5 gives the age in years

Youtube Rating Prediction

1) Search for videos "*Madonna*"

2) *Parse the json result*

3) *Extract features of videos, e.g. number of likes, number of dislikes, number of views, number of days since its publish date*

4) *Come up with a formula that relates features to the rating*

https://gdata.youtube.com/feeds/api/videos?q=madonna&max-results=50&v=2&alt=json

Example

- Try polynomial regression on youtube data and abalone examples
- How does the result of regression changes as we change the degree of polynomial from 5 to 2?
- Can you plot the in and out of sample R^2 score as a function of polynomial degree (k) for both problems? k=1..10

Group work

1) Extract the same set of financial indexes from 01/01/2000 to 01/01/2014
2) Pull the data for one the composite indexes (e.g. NASDAQ Composite .IXIC)
3) Try to come up with a linear or polynomial regression model that relates the indexes to the composite index
4) Assess the generality of your model by k-fold cross validation