Linear Regression and Model Selection Recitation

Nupur Chatterji
Kenny Marino
Colin White

Model Selection

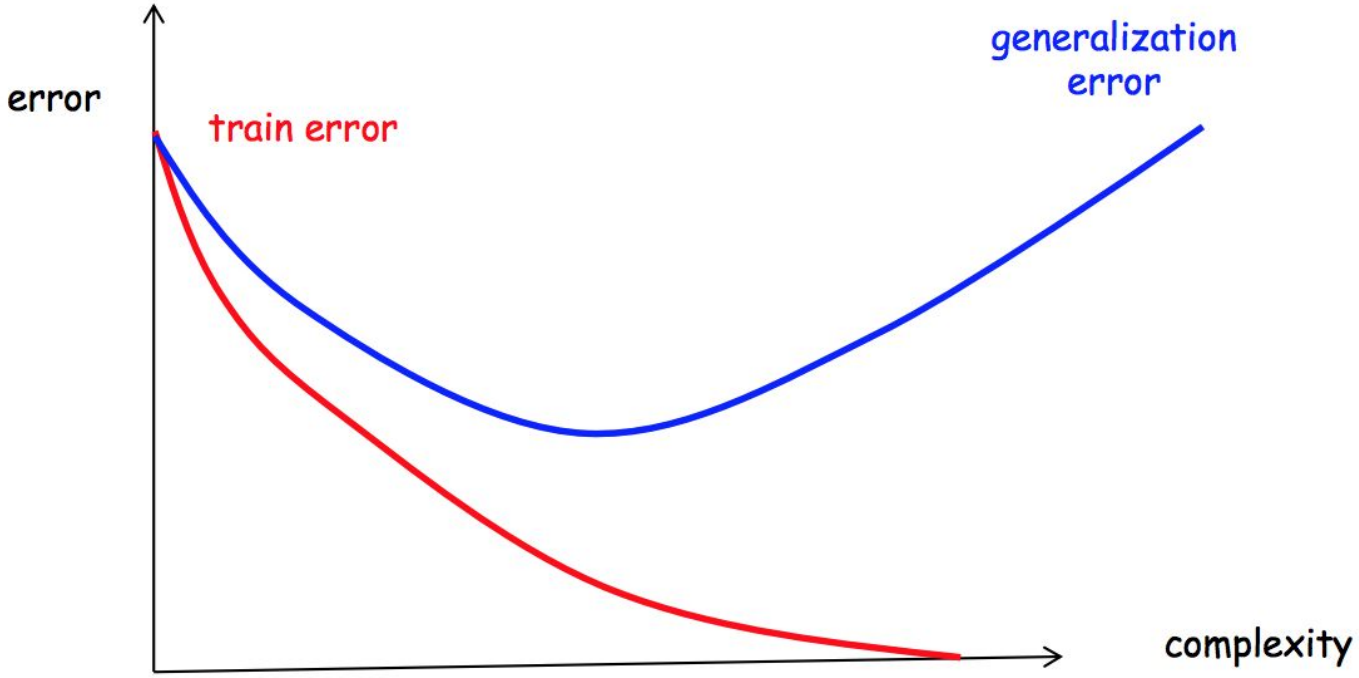
Aim: to find a hypothesis function that has the lowest error rate over the distribution

What does realizable mean?

What does agnostic mean?

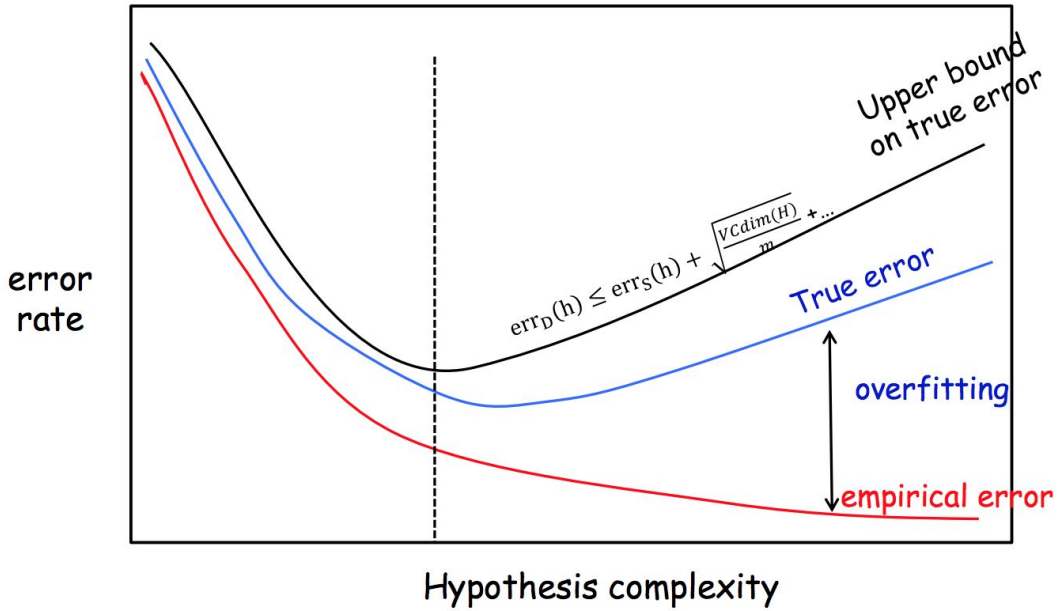
Note: we want the true error to be as low as possible, but we **CANNOT** measure that, so we bound the error of the distribution by the error of the sample

Model Selection



Structural Risk Minimization

$H_1 \subseteq H_2 \subseteq H_3 \subseteq \cdots \subseteq H_i \subseteq \ldots$



Linear Regression

Aim: to construct a predictor that minimizes error (based on how you choose to quantify error)

Univariate Case: fit a line of the form $f(x) = a + bx$ to the data

Multivariate Case: fit $f(x) = Xb$, where $X = [x_1, x_2, \ldots x_n]$ and $b = [b_1, b_2 \ldots b_n]^T$

Practice Questions for Linear Regression

3 Linear and Logistic Regression [20 pts. + 2 Extra Credit]

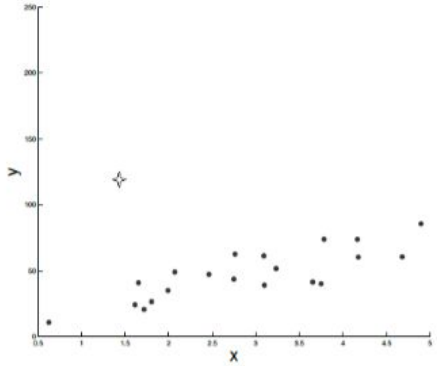
3.1 Linear regression

Given that we have an input x and we want to estimate an output y, in linear regression we assume the relationship between them is of the form $y = wx + b + \epsilon$, where w and b are real-valued parameters we estimate and ϵ represents the noise in the data. When the noise is Gaussian, maximizing the likelihood of a dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ to estimate the parameters w and b is equivalent to minimizing the squared error:

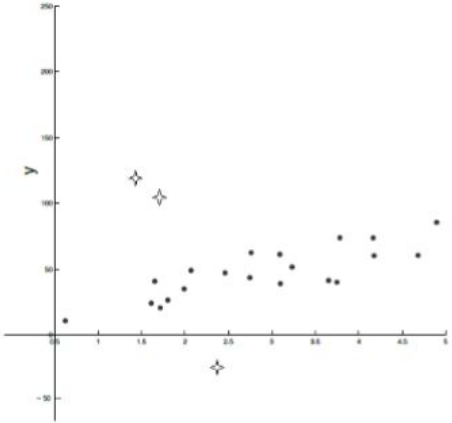
$$\arg\min_{w} \sum_{i=1}^{n} (y_i - (wx_i + b))^2.$$

Consider the dataset S plotted in Fig. 1 along with its associated regression line. For each of the altered data sets S^{new} plotted in Fig. 3, indicate which regression line (relative to the original one) in Fig. 2 corresponds to the regression line for the new data set. Write your answers in the table below.

Dataset	(a)	(b)	(c)	(d)	(e)
Regression line					



(a) Adding one outlier to the original data set.



(c) Adding three outliers to the original data set. Two on one side and one on the other side.

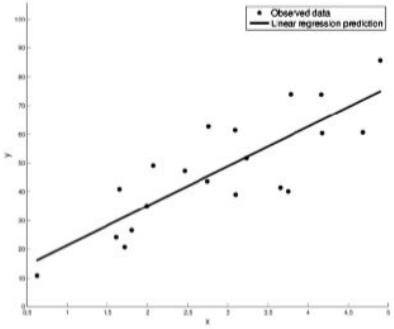
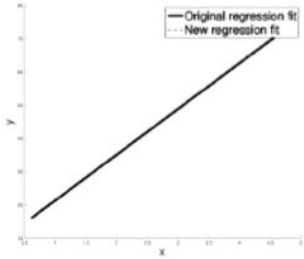
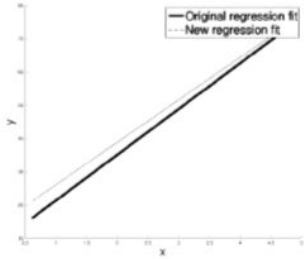


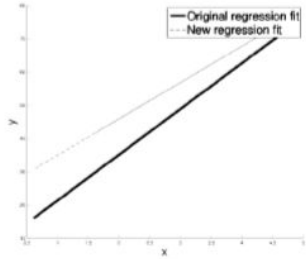
Figure 1: An observed data set and its associated regression line.



(a) Old and new regression lines.

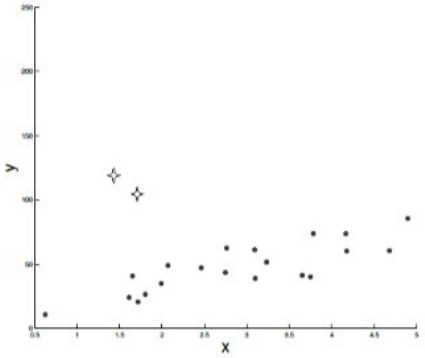


(b) Old and new regression lines.



(c) Old and new regression lines.

Figure 2: New regression lines for altered data sets S^{new}.



(b) Adding two outliers to the original data set.

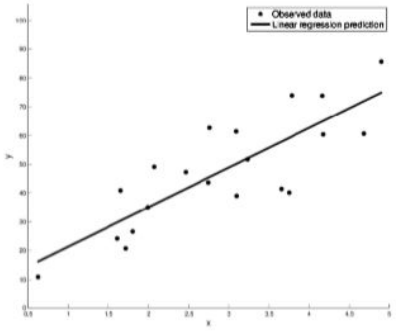
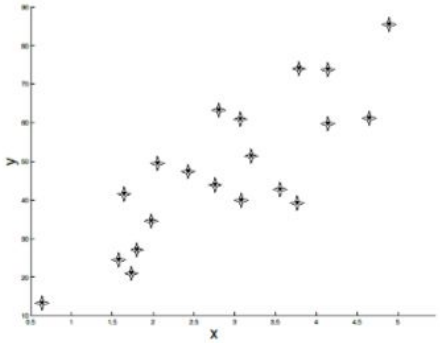
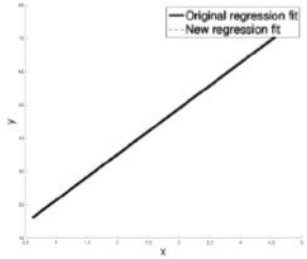


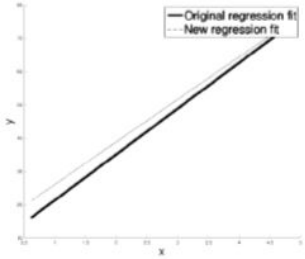
Figure 1: An observed data set and its associated regression line.



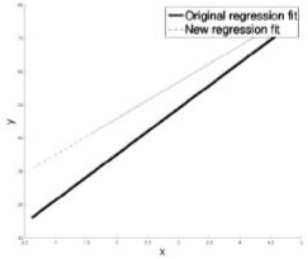
(d) Duplicating the original data set.



(a) Old and new regression lines.

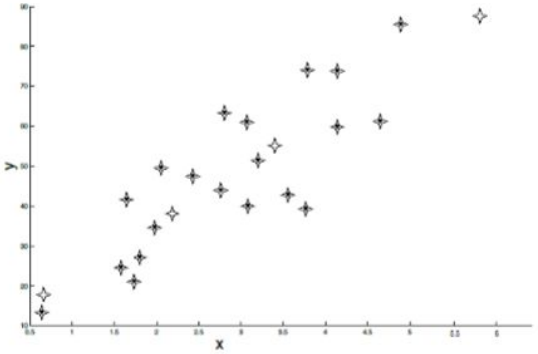


(b) Old and new regression lines.



(c) Old and new regression lines.

Figure 2: New regression lines for altered data sets S^{new}.



(e) Duplicating the original data set and adding four points that lie on the trajectory of the original regression line.

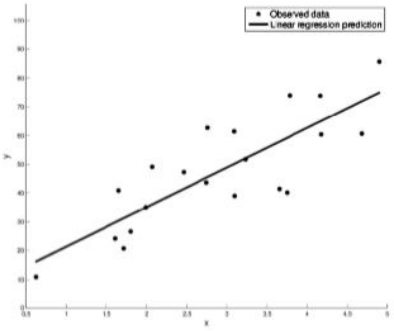
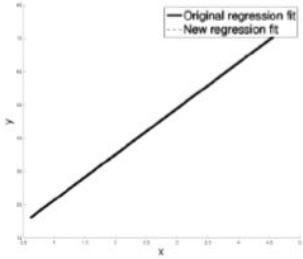
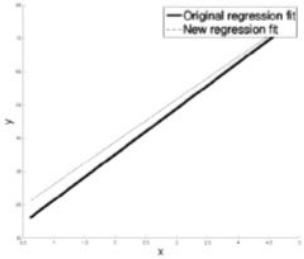


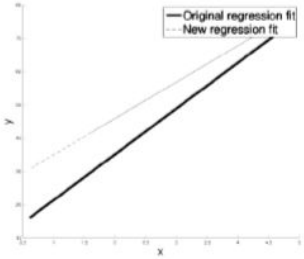
Figure 1: An observed data set and its associated regression line.



(a) Old and new regression lines.



(b) Old and new regression lines.



(c) Old and new regression lines.

Figure 2: New regression lines for altered data sets S^{new}.

4.1 Equivalence of maximizing conditional log-likelihood and minimizing squared-error loss. [6 pts.]

Given a dataset of inputs $x \in \mathbb{R}^d$ and real-valued outputs $y \in \mathbb{R}$, regression assumes each output y is a deterministic function f of input x, plus some zero-mean Gaussian noise ϵ:

$$y = f(x) + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{1}$$

This relationship means that y itself follows a Gaussian distribution $\mathcal{N}(f(x), \sigma^2)$.

Now consider learning the following specific function:

$$f(x_i) = w_0 + w_1 x_{i,1} + w_2 x_{i,2} \tag{2}$$

where $x_{i,j}$ denotes the j-th feature of the i-th example.

As discussed in class, calculating the parameter vector $w = [w_0, w_1, w_2]$ that maximizes the conditional data likelihood $\prod_i p(y_i|x_i, w)$ is equivalent to calculating the w that minimizes the sum of squared errors over the data. The derivation below proves this fact. In this derivation, **provide a short justification for why each line follows from the previous one:**

$$\hat{w} = \arg\max_w \prod_{i=1}^{n} p(y_i|x_i, w)$$

$$= \arg\max_w \ln \prod_{i=1}^{n} p(y_i|x_i, w)$$

$$= \arg\max_w \sum_{i=1}^{n} \ln p(y_i|x_i, w)$$

$$= \arg\min_w \sum_{i=1}^{n} -\ln p(y_i|x_i, w)$$

$$= \arg\min_w \sum_{i=1}^{n} -\left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\frac{(y_i - f(x_i))^2}{\sigma^2}\right)$$

$$= \arg\min_w \sum_{i=1}^{n} -\left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\frac{(y_i - (w_0 + w_1 x_{i,1} + w_2 x_{i,2}))^2}{\sigma^2}\right)$$

$$= \arg\min_w \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_{i,1} + w_2 x_{i,2}))^2$$

4.2 Other Forms of Regression. [5 pts.]

The above question shows that choosing regression parameters to maximize the conditional data likelihood is equivalent to choosing parameters that minimize the sum of squared errors for a particular function $f(x_i) = w_0 + w_1 x_{i,1} + w_2 x_{i,2}$. Here we explore whether this generalizes to other functions.

(a) [1 pt.] Suppose we wish instead to learn the following function:

$$f(x_i) = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + w_3 x_{i,3} \tag{4}$$

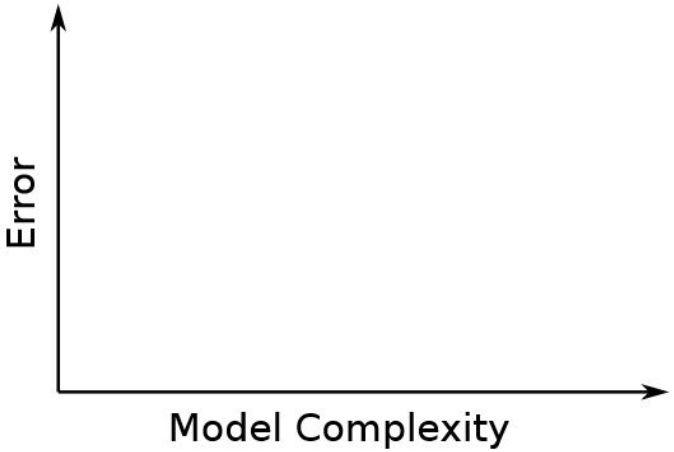
Can we derive a sum of squared objective corresponding to the maximum conditional likelihood estimate in this case? Answer by either giving the correct sum of squares objective to be optimized, or by explaining why we cannot do so in this case.

(b) [4 pt.] Suppose we wish instead to learn the following function, where y is no longer a linear function of x

$$f(x_i) = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + w_3 x_{i,1} x_{i,2} + w_4 x_{i,1}^2 \tag{5}$$

Can we derive a sum of squared objective corresponding to the maximum conditional likelihood estimate in this case? Answer by either giving the correct sum of squares objective to be optimized, or by explaining why we cannot do so in this case.

Practice with Model Complexity



On the graph,

- Curve of the training error?
- Curve of the test error?
- Optimal model complexity?
- Region of underfitting?
- Region of overfitting?

What happens when we double the dataset?