IEOR E4650 Business Analytics

Session 3: Analytical Framework

Spring 2018

Prof. Adam Elmachtoub

Today

Overview and framework for data analytics

- Types of Data
- Overview of methods
- Bias-Variance Tradeoff

"In God we trust, all others bring data." - William Edwards Deming

What is Data?

- *"Data is a set of values of qualitative or quantitative variables; restated, pieces of data are individual pieces of information." - Wikipedia*

- Anything recorded can be used as data, the quality of the analysis will be only as good as the quality of the data.

- Examples
  - Recorded information from sensors
  - Financial statements
  - Text records
  - Photos
  - Social Media

Examples of real data sources

Companies often have marketing, finance and operations data sets that can be used for analysis. Some additional methods of obtaining data:

- Mining client information
  - Who buys your products?
  - What do they think?

- Tracking online behavior
  - Who visits your websites?
  - How did they find it?
  - What do they do?
  - How long do they spend?

- Collaborative filtering
  - What do similar customers buy?
  - Ask customers to rate items
    (ex: recommendation at Amazon and netflix)

Data Collection

Data collection requires rigorous and systematic design and execution:

- thorough planning
- well-considered development
- appropriate management and analysis

Often the only way to obtain data is in the messy uncontrolled environments of the real-world

- Typically there will be a tradeoff between getting cleaner data from a controlled environment or more realistic but messier data in a real-world context.

Many companies naturally generate large amounts of data: Credit card companies, telecom companies, retail stores, etc.

Data Cleaning

Raw data need to be accurately entered for successful evaluation of information

- Check character variables have valid values
- Check numeric variables are within range
- Check for missing values
- Check for and eliminate duplicates
- Check for unique values (ID variables)
- Check for invalid dates or observations
- Combining multiple files
- Check for and eliminate duplicates

What Data Looks Like

	X_1	X_2	$\dots$	X_p	Y
x_1	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_2					
$\cdots$					
x_n					

- (x_i, y_i) is the i^{th} row of data, where x_i vector length p and $y_i \in \mathbb{R}$
- $X_1, X_2, \dots, X_p$ and Y are *each* scalar random variables, and we denote the random vector $X := (X_1, X_2, \dots, X_p)$
- One should think of x_i as a sample from X
- One should think of y_i as a sample from Y that depends on x_i
- X_i are also called input variables/independent variables/features/attributes/covariates/predictors/regressors/factors
- Y is also called output variable/outcome/response/label/dependent variable

Types of Variables

	X_1	X_2	$\dots$	X_p	Y
x_1	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_2					
$\cdots$					
x_n					

- Quantitative: Variables correspond to a natural numeric value such that close measurements are close in nature, e.g., variables such as weight, income, loyalty points
- Qualitative: Variables can be discretized or bucketed into categories, and there is no natural distance between categories. For example, ‘cat’ or ‘dog’, ‘0’ or ‘1’, $\{0, 1, 2, 3\}$, {January, February,..., December}. In some cases, categories have a natural order, i.e., ’18-35’, ’35-50’, ’50+’

Main Goal of Supervised Learning: Prediction

Explanatory/Input variables vs. Response/Output variables

1. Obtain some kind of model based on observations, or training data $\{(x_i, y_i), i = 1, 2, \ldots, n\}$, through a process called learning (or estimation).
2. Use that model to predict outputs based on new inputs from never-before-seen test data.

Types of Supervised Learning

- Regression: predict a quantitative/continuous output variable Y
- Classification: predict a qualitative/discrete/categorical output variable Y

Towards a final dataset

- Real data is often incomplete, with missing observation and values
- Sometimes variables will be more useful after a transformation. For example, we may want to look at a new variable log$(Wage)$ in addition to $Wage$.
- Sometimes variables will interact with each other, such as $Wage * Age$, and we create a new variable.
- A qualitative variable with K categories can be represented by $K - 1$ dummy binary variables. For example, dummy variable i is 1 if the category is i and 0 otherwise. If all dummy variables are 0, then the category is K

Big Data

The term *Big Data* is used to refer to data that either:

- has a large number of observations (n is large)
- has many more variables than observations ($p >> n$)
- is too large to be processed on a single machine

This raises some challenges.

- Cannot plot and check each variable or interactions between variables
- Choosing between many predicting variables, need to account for multiple hypothesis testing and watch out for false discovery
- Computational issues of processing very large amounts of data

Data Analytics

Data Analytics seeks to extract insights from raw data.
Related terms:

- *Machine Learning*
- *Data Mining*

Two stages:

- *Exploratory* - examine the data and come up with conjectures
  - Visualizing the data
  - Use statistical methods that will discover relations in the data
  - Variable selection and dimension reduction
- *Confirmatory* - test and validate conjectures
  - Hypothesis testing
  - Out of sample tests
  - Check for false discovery

Data Analytics Methods

- Supervised learning
  - Needs one or multiple outcome variables
  - Use some variables to predict unknown or future values of other variables
  - Some subproblems are prediction and classification
    ex: forecasting iPhone sales
- Unsupervised learning
  - Does not use outcome information
  - Finds interpretable patterns that describe the data
  - ex: Grouping students by characteristics

Data Analytics Methods and Goals

Data analytics are useful if they enable use to improve performance in the future. We measure of how good our solution is by whether :

- Regression, Classification - gives a good prediction for *new observations*
- Clustering - gives a useful organization of the data
- Optimization - finds the best possible solution
- Parameter Estimation - gives a precise estimate

In a business setting, a successful analytics solution should improve business performance. Often we will care more about getting interpretable results than about the raw prediction power. ex: We may want to use data to answer whether or not to continue an ad campaign

Data Analytics Methods

Regression - use data on past observation and their outcomes to predict outcome values for new observations
ex: life expectancy of patients, how a user will rate a movie

- Linear Regression, Subset Selection, Lasso Regressions, Regression Trees, K-Nearest Neighbors

Classification - use data of categorized observations to learn the categorize and classify new observations
ex: determine whether an email is spam, detect sick patients

- Logistic Regression, K-Nearest Neighbors, Decision Trees

Data Analytics Methods II

Clustering - find how to group observation
ex: find product categories, find communities in a social network

- K-Means Clustering, Hierarchical Clustering, Principal Component Analysis

Parameter Estimation - use data and a model to answer a specific question
ex: estimate the gravitational constant from experimental data, estimate consumer's sensitivity to an increase in price

- Bootstrap, Difference in Difference, Causal Inference, Shrinkage Estimators

Optimization - compute the option that maximizes an objective
ex: find the fastest delivery route, generate a financial portfolio that maximizes return for a given risk level

- Linear/Nonlinear, Integer Optimization, Sensitivity Analysis, Simulation Methods

Regression

Regression is the task of predicting quantitative response variables. The most common approach is linear regression, where we assumer there is a linear relationship between the continuous outcome variable Y and the explanatory variables $X_1, \ldots, X_p$, i.e.,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

The values of $\beta_0, \beta_1, \ldots, \beta_p$ are unknown but estimated by minimizing the mean squared error.

Given estimated values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ we can predict the value of new observations for which we know $X_1, \ldots, X_p$ but not Y.

Examples:

- Predicting sales based on advertising expenditures
- Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
- Time series prediction of macroeconomics and stock market indices

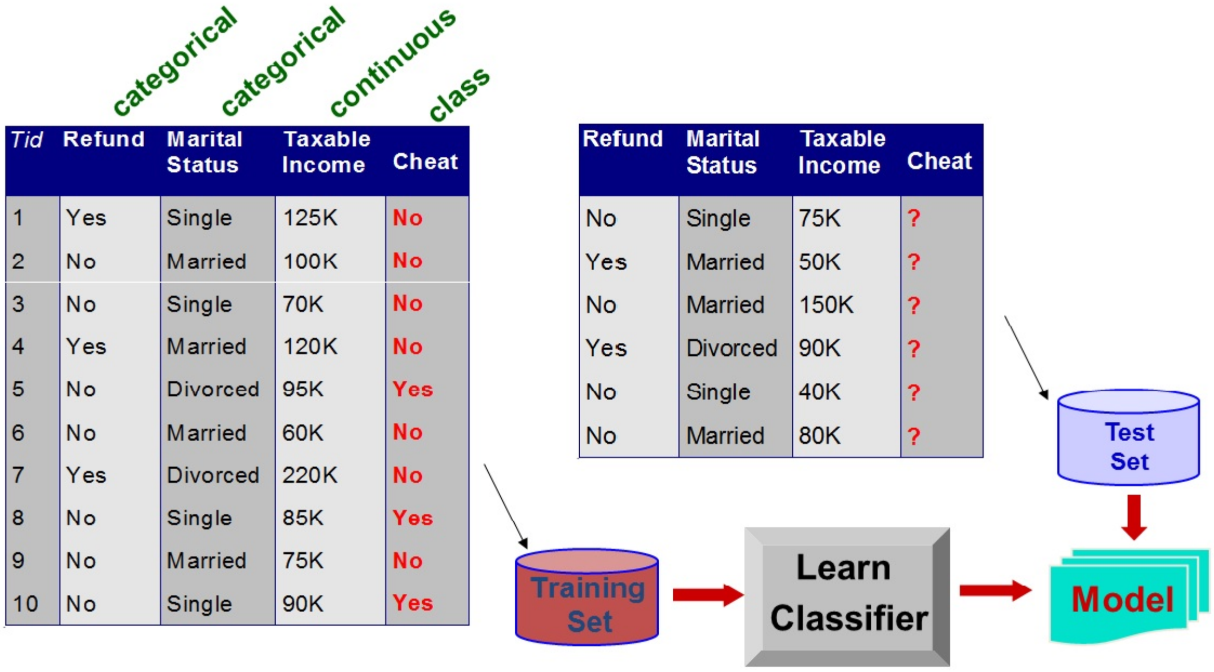
Classification 1: Credit Card Fraud

Goal: determine whether a credit card transactions is fraudulent.

Approach:

- Available variables: the credit card transaction and the information on its account-holder: When does a customer buy, what does he buy, how often he pays on time, etc
- Label past transactions as fraud or fair transactions. This forms the *class* labels.
- Learn a model for the class of the transactions
- Use this model to classify new transaction, and warn if the classifier determines a transaction to be fraud
- Monitor and check performance

Classification 2: Tax Return



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

Train and Test

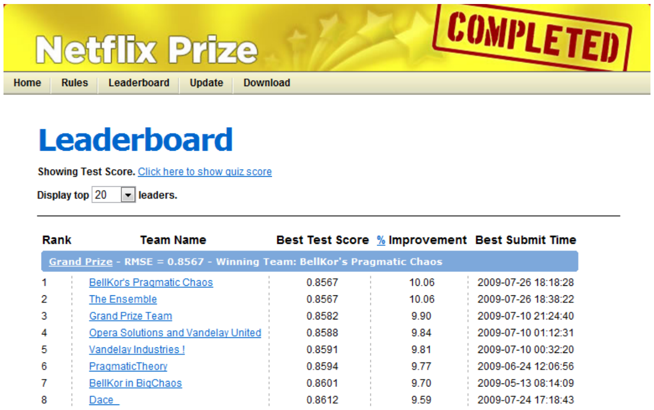
We want to use the available data to learn how good our forecast will be for new observations. Statistical measures such as R^2 give in-sample fit, not out-of-sample fit.

If the data is big enough, divide the data into a training set and a test set. Estimate using the training set, and use the test set to simulate new data.

Example: The Netflix Prize

Netflix asked participants to submit algorithm for predicting user ratings for films. A training data set provided by Netflix included about 100m ratings that 480k users gave to 18k movies. Each entry include a user ID, movie ID, date, and rating.

Algorithm were tested by asking to predict ratings on a different test set, for which ratings were hidden. Scores were calculated according to how predicted ratings compared with the (hidden) actual ratings.



Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 leaders.

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43

Clustering

Given a set of data points, each having a set of attributes, and a similarity measure among them, find *clusters* such that

- Data points in one cluster are more similar to one another
- Data points in separate clusters are less similar to one another

Similarity depends on context and usage. We may want to cluster movie viewers so clusters will include viewers with similar taste, or we may want to cluster movie viewers according to their geographical location.

Model Flexibility

Let us return to the prediction problem, and suppose we wish to predict outcome Y using one variable X. Our predictors will be of the form $Y = f(X) + \epsilon$ for various function $f(\cdot)$.
For example:

- $f(x) = \beta_0$, with β_0 estimated from the data
- $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$, with $\beta_0, \beta_1, \beta_2$ estimated from the data
- $f(x) = \sum_{k=0}^{10} \beta_k x^k$ with $\{\beta_k\}_{k=0\ldots10}$ estimated from the data

Changing the function/model impacts the *bias* and *variance*.

- The *bias* refers to the model's ability, on average, to closely predict the response variable in the training dataset
- The *variance* refers to the model's stability, or how much the predictions would change if we had different training datasets

Model Flexibility and Bias Variance Tradeoff

The is typically a tradeoff between bias and variance:

- A very flexible model will result in lower bias on the training data, but will typically have higher variance across different training datasets.
- A very inflexible model will result in higher bias on the training data, but will typically have smaller variance across different training datasets.

To get the best prediction out-of-sample we need to balance the two sources of error. Model selection is the process of tuning the model to achieve the best bias-variance tradeoff. Of course, to select a model, we need to be able to measure its performance. This is know as model assessment.

We will come back to these central ideas many times during the course.

Bias-Variance Tradeoff: Examples

Data points $\mathcal{D} = \{(x_i, y_i)\}_{i=1 \dots n}$ and we fit three curves to the data:

- The orange curve is an estimated linear regression line $\hat{f}(x) = \beta_0 + \beta_1 x$
- The blue and green curves estimate $\hat{f}$ using splines (higher order polynomials)

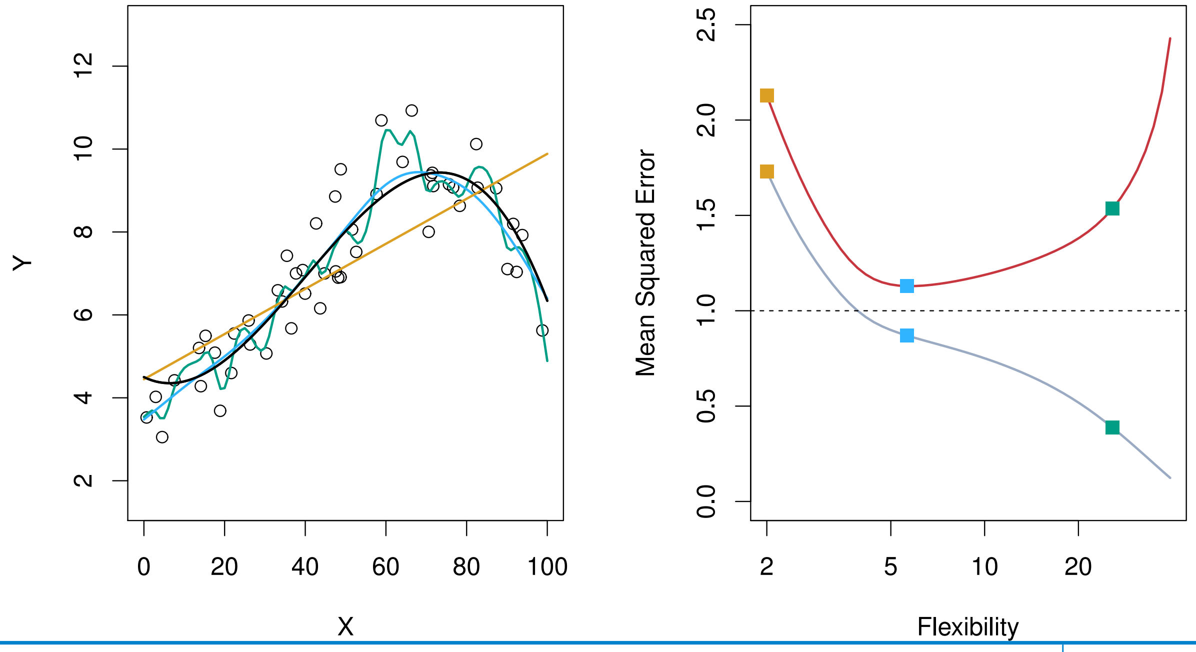
We will measure how well our fitted curves fit in-sample by calculating the *mean squared error* (MSE) over our training data $\{(x_i, y_i)\}_{i=1 \dots n}$

$$\text{MSE}_{\mathcal{D}}(\hat{f}) = \frac{1}{n} \sum_{(x_i, y_i) \in \mathcal{D}} (y_i - \hat{f}(x_i))^2$$

And we will check how well the curves fit out-of-sample by calculating MSE on new test data $\mathcal{D}' = \{(x_j, y_j)\}_{j=1 \dots m}$.

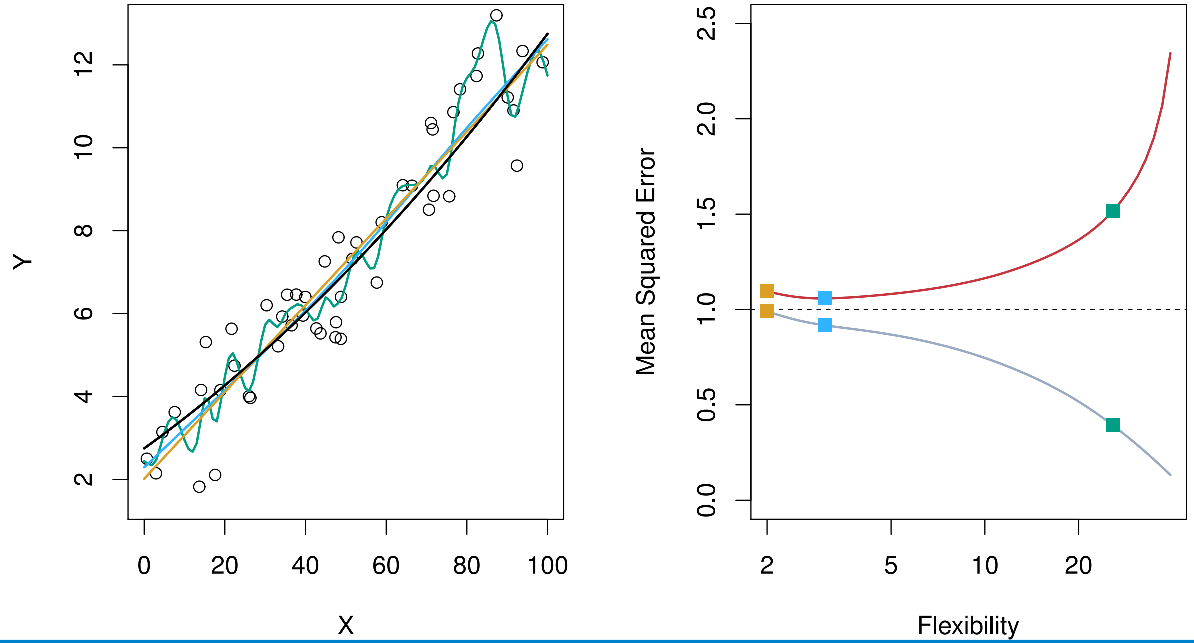
Bias-Variance Tradeoff: Examples

The plot on the left shows the points (x, y) and three fitted curves. The plot on the right shows the mean squared errors of each curve for the training data (grey curve) and test data (red curve).



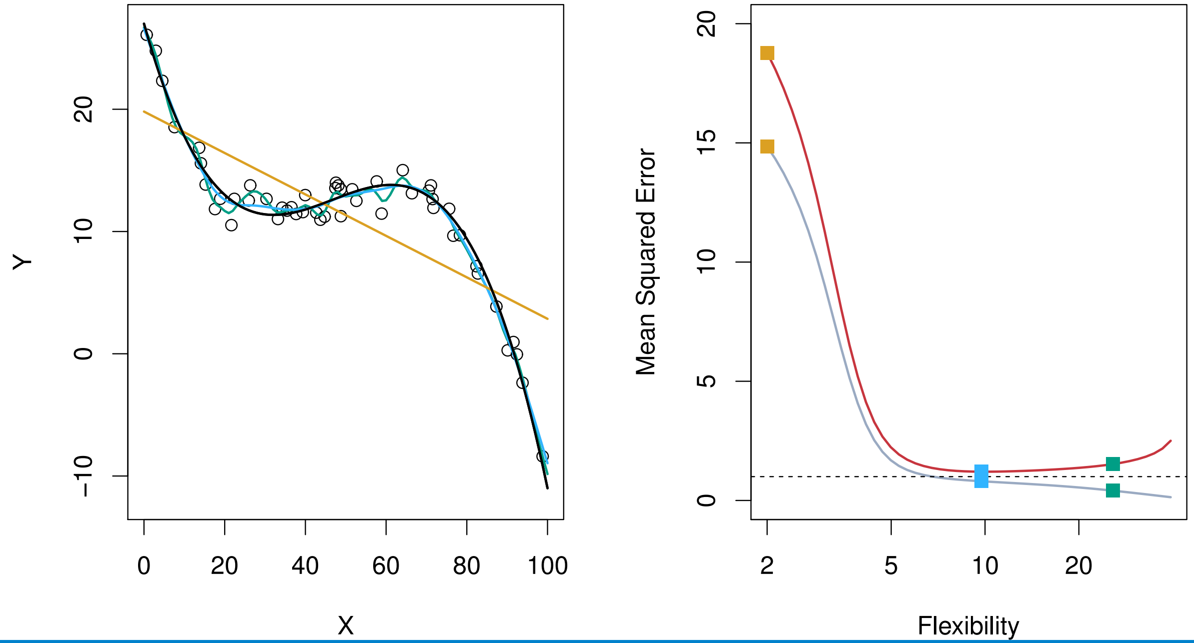
Bias-Variance Tradeoff: Examples

The plot on the left shows the points (x, y) and three fitted curves. The plot on the right shows the mean squared errors of each curve for the training data (grey curve) and test data (red curve).



Bias-Variance Tradeoff: Examples

The plot on the left shows the points (x, y) and three fitted curves. The plot on the right shows the mean squared errors of each curve for the training data (grey curve) and test data (red curve).



Model Selection: Bias-Variance Tradeoff

Theorem.

$$\mathcal{E}_{test} = E_{D\sim\mathcal{D}}\left[E_{x\sim X}\left[(g^{D}(x) - \bar{g}(x))^2\right]\right] + E_{x\sim X}\left[(\bar{g}(x) - f(x))^2\right] + \sigma^2$$

- The first summand is the variance of $g^{\mathcal{D}}(x)$ around $\bar{g}(x)$.
  - $g^{D}(\cdot)$ is estimate of function $f(\cdot)$ from our training sample D which itself is randomly sampled from $\mathcal{D}$
  - $g^{D}(x)$ is estimate of $f(x)$ from our training sample D
  - $\bar{g}(\cdot)$ is expectation of $g^{D}(\cdot)$ over all training samples $\mathcal{D}$
  - $\bar{g}(x)$ is average estimate for a specific input x
- The second summand is the squared of the bias $\bar{g}(x) - f(x)$.
- The third summand is known as the irreducible error, i.e., the variance of the actual noise ϵ

- A very flexible model results in low squared bias but high variance (overfitting)
- A very restricted model in high squared bias and low variance (underfitting)