Descriptive Statistics

Steve Avsec

Illinois Institute of Technology

August 19, 2025

Overview

A Definition

Statistical Learning is a set of tools for understanding data using statistical methods.

A Definition

Statistical Learning is a set of tools for understanding data using statistical methods.

Three levels:

1. *Descriptive* statistics are a set of tools for describing a static data set.

A Definition

Statistical Learning is a set of tools for understanding data using statistical methods.

Three levels:

1. *Descriptive* statistics are a set of tools for describing a static data set.
2. *Predictive* learning is a set of tools for predicting an outcome from historical data.

A Definition

Statistical Learning is a set of tools for understanding data using statistical methods.

Three levels:

1. *Descriptive* statistics are a set of tools for describing a static data set.
2. *Predictive* learning is a set of tools for predicting an outcome from historical data.
3. *Prescriptive* analysis is a set of tools for prescribing a business action from data.

Types of Variables

1. *Continuous* variables can take any value in a specified interval (e.g. $(0, 1)$ or $(-\infty, \infty)$).

Types of Variables

1. *Continuous* variables can take any value in a specified interval (e.g. $(0, 1)$ or $(-\infty, \infty)$).

2. *Categorical* variables can take any value in a finite set (e.g. $\{A, B, C\}$ or $\{0, 1\}$ or the set of all states in the U.S.).

Types of Variables

1. *Continuous* variables can take any value in a specified interval (e.g. $(0, 1)$ or $(-\infty, \infty)$).
2. *Categorical* variables can take any value in a finite set (e.g. $\{A, B, C\}$ or $\{0, 1\}$ or the set of all states in the U.S.).
3. *Ordinal* variables can take values in an ordered set (almost alway finite). E.g. how would you rate your pain on a scale of 1-10? How would you rate your service today?

Supervised vs. Unsupervised Learning

- *Supervised* Learning is predictive with a set of historical data with given outcomes for each case.

Supervised vs. Unsupervised Learning

- *Supervised* Learning is predictive with a set of historical data with given outcomes for each case.
    - A *regression* problem is when the outcome is a continuous variable.

Supervised vs. Unsupervised Learning

- *Supervised* Learning is predictive with a set of historical data with given outcomes for each case.
  - A *regression* problem is when the outcome is a continuous variable.
  - A *classification* problem is when the outcome is a categorical variable.

Supervised vs. Unsupervised Learning

- *Supervised* Learning is predictive with a set of historical data with given outcomes for each case.
  - A *regression* problem is when the outcome is a continuous variable.
  - A *classification* problem is when the outcome is a categorical variable.
  - A *ordinal regression* problem is when the outcome is ordinal.

Supervised vs. Unsupervised Learning

- *Supervised* Learning is predictive with a set of historical data with given outcomes for each case.
  - A *regression* problem is when the outcome is a continuous variable.
  - A *classification* problem is when the outcome is a categorical variable.
  - A *ordinal regression* problem is when the outcome is ordinal.
- *Unsupervised* Learning is a set of techniques that can be either descriptive or predictive but have no outcomes attached.

Supervised vs. Unsupervised Learning

- *Supervised* Learning is predictive with a set of historical data with given outcomes for each case.
  - A *regression* problem is when the outcome is a continuous variable.
  - A *classification* problem is when the outcome is a categorical variable.
  - A *ordinal regression* problem is when the outcome is ordinal.
- *Unsupervised* Learning is a set of techniques that can be either descriptive or predictive but have no outcomes attached.
- *Features* or *Covariates* are variables in data sets that are not outcomes.

Basic Setup

Let Y be the output (also called a *response* or *target variable*).

Basic Setup

Let Y be the output (also called a *response* or *target variable*).

Let $X = (X_1, \ldots, X_d)$ be a vector of features (or covariates).

Basic Setup

Let Y be the output (also called a *response* or *target variable*).

Let $X = (X_1, \ldots, X_d)$ be a vector of features (or covariates).

Find a (computable mathematical) function f such that

$$Y = f(X) + \varepsilon$$

Parametric vs. Nonparametric

Suppose f takes the form:

$$f(X) = \beta_0 + \beta_1 * X_1 + \ldots + \beta_d * X_d$$

Parametric vs. Nonparametric

Suppose f takes the form:

$$f(X) = \beta_0 + \beta_1 * X_1 + \ldots + \beta_d * X_d$$

(This is just linear regression.)

Parametric vs. Nonparametric

Suppose f takes the form:

$$f(X) = \beta_0 + \beta_1 * X_1 + \ldots + \beta_d * X_d$$

(This is just linear regression.)

A *parametric* model is a model where f is chosen from a parameterized set of functions.

Parametric vs. Nonparametric

Suppose f takes the form:

$$f(X) = \beta_0 + \beta_1 * X_1 + \ldots + \beta_d * X_d$$

(This is just linear regression.)

A *parametric* model is a model where f is chosen from a parameterized set of functions.

A *nonparametric* model is a model where f is estimated directly from the data without a general closed form expression.

Accuracy versus Interpretability

Consider a fancier model:

$$f(X) = \sum_{\alpha | |\alpha| < n} \beta_\alpha * X^\alpha$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multiindex and $|\alpha| = \alpha_1 + \cdots + \alpha_d$. (E.g. f is a polynomial of degree n).

Accuracy versus Interpretability

Consider a fancier model:

$$f(X) = \sum_{\alpha||\alpha|<n} \beta_\alpha * X^\alpha$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multiindex and $|\alpha| = \alpha_1 + \cdots + \alpha_d$. (E.g. f is a polynomial of degree n).
This is a much more *flexible* model than linear regression, but also has many more parameteters.

Accuracy versus Interpretability

Consider a fancier model:

$$f(X) = \sum_{\alpha||\alpha|<n} \beta_\alpha * X^\alpha$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multiindex and $|\alpha| = \alpha_1 + \cdots + \alpha_d$. (E.g. f is a polynomial of degree n).
This is a much more *flexible* model than linear regression, but also has many more parameteters.
Polynomials aren't usually well-suited for most modeling tasks but many supervised models use many more parameters (GAMs, trees, SVMs, neural networks).

Accuracy versus Interpretability

Consider a fancier model:

$$f(X) = \sum_{\alpha \mid |\alpha| < n} \beta_\alpha * X^\alpha$$

where $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multiindex and $|\alpha| = \alpha_1 + \cdots + \alpha_d$. (E.g. f is a polynomial of degree n).
This is a much more *flexible* model than linear regression, but also has many more parameteters.
Polynomials aren't usually well-suited for most modeling tasks but many supervised models use many more parameters (GAMs, trees, SVMs, neural networks).
High interpretability: fewer parameters, clearer relationships between input/output. High accuracy: more parameters, tighter fitting functions

Metrics

Mean squared error:

$$MSE = \frac{1}{n}\sum_{j=1}^{N}(y_j - f(\mathbf{x}_j))^2$$

(Very widely used as a goodness-of-fit metric for regression problems.)

Metrics

Mean squared error:

$$MSE = \frac{1}{n}\sum_{j=1}^{N}(y_j - f(\mathbf{x}_j))^2$$

(Very widely used as a goodness-of-fit metric for regression problems.)

Precision:

$$P = \frac{TP}{TP + FP}$$

Recall:

$$R = \frac{TP}{TP + FN}$$

Bias versus Variance

For MSE:

$$E(y_0 - f(\mathbf{x}_0))^2 = Var(f(\mathbf{x}_0)) + Bias(f(\mathbf{x}_0)) + Var(\varepsilon)$$

Bias versus Variance

For MSE:

$$E(y_0 - f(\mathbf{x}_0))^2 = Var(f(\mathbf{x}_0)) + Bias(f(\mathbf{x}_0)) + Var(\varepsilon)$$

The first term is the variance introduced by changing the *training* set. If $f(\mathbf{x}_0)$ changes by large amounts by taking different samples of training data, it's high variance (usually a more flexible model).