Machine Learning Concepts

CEU, Winter 2024

Instructors: Robert Lieli and Janos Divenyi

What is machine learning (ML)?

In a narrow sense:

- A set of modern **statistical methods** designed to handle complex and high dimensional prediction and classification tasks
- E.g., lasso and ridge regression, logistic regression, regression trees, random forest, neural networks, etc.
- But more traditional statistical estimation methods (such as linear regression or kernel regression) can also be labeled as machine learning methods

What is machine learning?

In a broader sense:

- A form of inductive reasoning — based on (X, Y) examples in the available in the *training* data, *learn* the relationship between X and Y
  - What is the Y value associated with a newly observed X value?

  $\Rightarrow$ **supervised learning**

- Finding stable patterns in the data without examples (no observed Y values)
  - E.g., how can we compress the information contained in a large X vector into one or two variables?

  $\Rightarrow$ **unsupervised learning**

Why study machine learning?

- We are in the “big data” era
- Large data sets collected by apps, websites, government institutions (administrative data), etc.
  - Large means two things: i) many observations; ii) many variables
  - Modern ML methods are particularly helpful in prediction tasks when there are many variables compared to the sample size
- The emergence of **data science** as its own field. It melds:
  - statistics/mathematics
  - computer science
  - the various domains of application: engineering, medicine, economics, etc.

Applications of Machine Learning

Machine learning works. Amazing modern applications:

- Great advances toward self-driving cars
- Speech/face/image recognition
- Chess engines that easily beat the best human players
- Personalized ads, recommendations, spam filer, etc.
- Large Language Models, Generative Artificial Intelligence

We will restrict ourselves to simpler, less glamorous prediction problems

Formalizing prediction problems (ISLR Ch. 2)

- Y = outcome of interest (dependent variable, output, etc.)
- X = vector of predictors (covariates, features, independent variables, inputs, etc.):

  $$X = (X_1, X_2, \ldots, X_p)'$$

  where p can be large.

Examples

1. Y=price of an AirBnB apartment

   X = location, size, amenities, etc.

2. Y = baby's birthweight

   X = mother's age, medical history, prenatal care utilization, socio-economic status, zip code, smoking status, etc.

Examples

3. $Y = 1$ if an individual suffers a heart attack within the next five years; $Y = 0$ otherwise

   $X =$ blood pressure, cholesterol level, age, gender, smoking status, diabetes, physical activity, etc.

4. $Y =$handwritten digit $\in \{0, 1, \ldots, 9\}$

   $X =$ frame digit in a square;
   break down square into pixels;
   define a dummy variable for each pixel that shows whether the pixel is marked or empty

Examples 3 and 4 are classification problems (Y is binary/discrete)

This course: emphasis prediction on problems with continuous Y (but some methods can handle both)

Formalizing prediction problems

Our basic model of the data generating process is:

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

where

- ϵ is a random noise term independent of X;
- the function $f(\cdot)$ captures the systematic relationship between X and Y
- In fact, because $X \perp \epsilon$,

  $$f(X) = E(Y|X),$$

  i.e., $f(X)$ is the conditional mean (expectation) of Y given X.
- $f(X) = E(Y|X)$ is one's "best" guess of Y given X. Best = **smallest mean squared error**.

What machine learning methods do

- Let $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ be a sample of observations on (X, Y) — the **training sample**
- The goal of statistical/machine learning is to construct

  an **estimate** $\hat{f}(x)$ of the function $f(x)$

  from the training sample
- In other words, we want to "learn" about the systematic relationship between X and Y from the data

Example

- Let $X = (X_1, X_2, X_3)$
- $Y = X_1 + X_1X_2 + X_2^2 + \epsilon$
  - $f(X) = X_1 + X_1X_2 + X_2^2$
    $\Rightarrow$ true systematic relationship between X and Y

We could try to model the relationship between X and Y as:

- $\hat{f}_1(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$,
  where the $\hat{\beta}$ coefficients are estimated by OLS
- $\hat{f}_2(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 + \hat{\beta}_4 x_1 x_2$,
  where the $\hat{\beta}$ coefficients are estimated by OLS
- $\hat{f}_3(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 + \hat{\beta}_4 x_1 x_2 + \hat{\beta}_5 x_3 + \hat{\beta}_6 x_3^2 + \hat{\beta}_7 x_1 x_3 + \hat{\beta}_8 x_2 x_3$,
  where the $\hat{\beta}$ coefficients are estimated by lasso

How do we evaluate predictive models?

- General principle: $\hat{f}(X)$ should provide an accurate prediction of Y on **new (test) data**
- Let X_{n+1} be an additional independent observation on X. We form the prediction $\hat{Y}_{n+1} = \hat{f}(X_{n+1})$
- We want the **prediction error**

  $$Y_{n+1} - \hat{Y}_{n+1} = Y_{n+1} - \hat{f}(X_{n+1})$$

  to be small for:

  i) some specific value x_0 of X_{n+1};
  ii) or, on average, over all possible values of X_{n+1}.
- Equivalently, we want $\hat{f}(x_0)$ to be close to $f(x_0)$ for:

  i) some specific value x_0 of X_{n+1};
  ii) or, on average, over all possible values of X_{n+1}.

What constitutes a good prediction in theory

Suppose that we are interested in obtaining accurate predictions for $X_{n+1} = x_0$.

That is, consider $\hat{Y}_{n+1} = \hat{f}(x_0)$ and $Y_{n+1} = f(x_0) + \epsilon_{n+1}$.

The theoretical **mean** (expected) **squared prediction error** (MSPE) at x_0 is defined as:

$$
\begin{aligned}
MSPE(x_0) &= E\left[(Y_{n+1} - \hat{f}(x_0))^2\right] \\
&= E\left[(\hat{f}(x_0) - f(x_0))^2\right] + Var(\epsilon_{n+1}) \\
&= MSE[\hat{f}(x_0)] + Var(\epsilon_{n+1}).
\end{aligned}
$$

- $MSE[\hat{f}(x_0)]$=the mean squared error of the **model** at $X = x_0$
- $Var(\epsilon_{n+1})$=irreducible prediction error.

Decomposing the MSE of the model

We can write the MSE of the model as the sum of two components:

$$
\begin{aligned}
MSE[\hat{f}(x_0)] &= E\left[(\hat{f}(x_0) - f(x_0))^2\right] \\
&= \left\{E[\hat{f}(x_0)] - f(x_0)\right\}^2 + E\left\{\hat{f}(x_0) - E[\hat{f}(x_0)]\right\}^2 \\
&= \{bias[\hat{f}(x_0)]\}^2 + Var[\hat{f}(x_0)]
\end{aligned}
$$

Meaning of bias and variance

- **Bias** = the difference between the average prediction, computed over all possible training samples, and the optimal prediction;
- **Variance** = how much the prediction varies from training sample to training sample around the average prediction.
- There is a fundamental tradeoff between these two quantities as one changes the flexibility/complexity of the predictive model $\hat{f}(x)$

How can we compute the mean squared (prediction) error for different models and learning methods?

1. Analytical calculations
2. Monte Carlo (computer) simulations in specific hypothetical scenarios
3. We can estimate the average *MSPE* over the possible values of X from an independent **test/validation sample**