University of Wisconsin - Madison
Department of Industrial and Systems Engineering
ISyE 521: Machine Learning in Action
Fall 2022

Lecture 0: Introduction to machine learning

Instructor: Justin J. Boutilier

Sept 7, 2022

In this lecture, we will introduce the field of machine learning using a real-world example. We will learn:

1. The difference between machine learning and artificial intelligence
2. The language of machine learning
3. The basics of linear regression

Aside: This course will not cover the (fascinating) history of machine learning. For those that are curious, this video provides a glimpse into the life and career of the *"Godfather of Deep Learning"*, Prof. Geoffrey Hinton.

*Corrections provided by Eric Pullick, Codanda Monappa, Jin-ri Lee, and Sofia Noejovich.

Example: Predicting the quality of wine

In March 1990, Princeton Economics Professor, Orley Ashenfelter, announced that he could predict the quality of *Bordeaux wines* without tasting them. *Bordeaux wine* is exclusively produced in the Bordeaux region of France and although it has been produced in the same way for hundreds of years, there can be significant differences in quality from vintage (i.e., year) to vintage. Bordeaux wine is believed to taste better with age, but it's hard to predict the future quality of wine when it is tasted at a "young" age. After Ashenfelter's announcement, Robert Parker, the world's influential wine expert, said:

"Ashenfelter is an absolute total sham...like a movie critic who never goes to see the movie but tells you how good it is based on the actors and director...ludicrous and absurd"

So the question is:

Can machine learning beat a human expert at predicting wine quality?

The language of machine learning

Before we can answer that question (and many others!), we first need to define the language of machine learning.

What is machine learning?

- Simply put, the field of machine learning focuses on teaching machines how to learn.
- As Prof. Tom Mitchell put it: *"Machine learning is the study of computer algorithms that improve automatically through experience."*
- Note that machine learning algorithms are not provided with explicit instructions on how to perform a given task. Instead, the algorithms rely on patterns and experience to effectively learn how to perform the task.

How is machine learning different from artificial intelligence?

- Artificial intelligence focuses on teaching computers *"to behave in ways that, until recently, we thought required human intelligence."* (Prof. Andrew Moore)
- In other words, AI is a moving target, based on what tasks we currently believe require human intelligence
  - For example, the first calculator (or first computer) could be considered as an early form of artificial intelligence
  - Why? Because, at the time, we thought that complex computations required human intelligence. In fact, early computers were not trusted to be accurate and humans were needed to check their computations (see the 2016 movie Hidden Figures for a popular example)
- Machine learning is one (of potentially many) ways to achieve artificial intelligence

Definitions

- *Target:* The outcome that we want to predict. Often referred to as the *dependent variable*.
- *Features:* The information that we use to predict the target. Often referred to as *independent variables*.
- *Observations:* The data sample that we observe, including features and (sometimes) targets. Often referred to as the *dataset*.
- *Model:* A function that maps the features to the target. Sometimes referred to as the *machine*.
- *Hyperparameters:* Tunable model parameters.
- *Training and testing*: A process that partitions the observations into disjoint sets (i.e., they have no elements in common):
  - *Training set:* The observations that we use to train the model to perform a task.

  - *Validation set:* The observations that we use to validate our modelling (hyper-parameter) choices.
  - *Testing set:* The observations that we use to test the model performance using data that has not been seen before.

- *Fit:* How well the model predicts the target.

How do machines learn?

1. *Supervised learning:* teaches the machine to perform a task using observations that include both features and targets

2. *Unsupervised learning:* teaches the machine to perform a task using observations that only include features (no targets).

3. *Reinforcement learning:* teaches the machine to perform a task without using prior observations. Instead, we teach the machine to learn through a process of trial and error.

4. *Deep learning:* teaches the machine to learn in an abstract way, without a specific task in mind. Deep learning can be both supervised and unsupervised.

Classification vs. Regression

Machine learning problems can be classified into two categories based on the type of target:

1. *Regression:* The target variable is continuous.

2. *Classification:* The target variable is discrete. In the simplest case, the target is binary corresponding to two groups (e.g., dead or alive, win or loss), but multi-class problems can include a large number of groups (e.g., multiple modes of transportation for commuting).

Example continued: Predicting the quality of wine

The price of a wine is often used as a proxy for the quality (high quality wines have higher prices than low quality wines). Ashenfelter focused on predicting the price of wine and discovered two explanations for the observed variations in price:

- Age: older wines are more expensive

- Weather: varies dramatically from one year to the next in Bordeaux

Based on his exploratory data analysis, Ashenfelter decided to use a linear regression model to predict the price of wine.

What features were used to predict the price of wine?

Ashenfelter used four features:

- Age
- Winter rain (October - March)
- Harvest rain (August - September)
- Average growing season temperature (April - September)

How was the price of wine defined?

- The price index measures the price of many different wineries (i.e., wines) in thousands of wine auctions during a given year. It is normalized to the highest price in the dataset.
- Ashenfelter used log(Price index) as the target in his linear regression model.
- Why log(Price index)?
  - Produces a better linear fit (discovered through plotting)
  - Inflation and other factors typically increase prices exponentially, so a log transformation generates a more linear price trend

The final model

$$Log(Price\ Index) = -12.145 + 0.001167 * WinterRain + 0.616 * AvgGrowingSeasonTemp + 0.0238 * Age - 0.00286 * HarvestRain$$

The model predicts wine price at maturity, which reflects the wine's true quality. Based on the model, *Wine* magazine stated:

"the formula's self-evident silliness invites disrespect"

while the Bordeaux wine industry was outraged that their craft could be reduced to a simple equation:

"How can he know anything about wine that he hasn't tasted?...Would you listen to a food critic who has never tasted the food?"

How did the model perform?

The model had an $R^2 = 0.83$ (meaning that the model explains 83% of the variation in price) and all features were found to be statistically significant. Further testing revealed that the predictions were robust to the addition of other features.

- Parker (wine expert) in 1991:

  - 1986 is *"very good to sometimes exceptional"*
  - Other years are average

- Ashenfelter (modeler) in 1991:
  - 1986 is mediocre
  - 1989 will be *"the wine of the century"* and 1990 will be even better!

- In 2003, there was virtually unanimous agreement that 1989 and 1990 are outstanding wines
  - 1989 sold for more than twice the price of 1986 and 1990 sold for even higher prices!
  - *"Two of the best vintages in the last 50 years"*

- In 2003, Ashenfelter predicted 2000 and 2003 would be equally great
  - Parker has stated: *"2000 is the greatest vintage Bordeaux has ever produced"*

In summary, we've learned that a linear regression model with only a few variables can accurately predict wine prices. The model is objective, rational, emotionless, and unbiased (in principle). In many cases, the model outperforms wine experts' opinions and provides a quantitative approach to a traditionally qualitative problem. You can find Ashenfelter's original paper here or here. For more information and ongoing wine analysis, see here.

Method: Linear regression

Linear regression is one of the simplest and most widely used prediction methods. It is used to model the relationship between a dependent variable, called the *target variable* and one or more independent variables, called the *feature variables*. For a given data point, let y represent the target variable and $x_1, x_2, \ldots, x_F$ represent F feature variables. Our aim is to determine the best values for the regression coefficients, $\beta_0, \beta_1, \ldots, \beta_F$ such that our predicted target values:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_F x_F$$

are as close as possible to the true target values (denoted y). To do this, we define the error (or residual) for a given observation to be $(\hat{y} - y)$. Then, we choose our regression coefficients to minimize the mean squared error (MSE) between the true target values and the predicted target values across all data points. For example, suppose we observe n data points of the form $(y_i, x_{i1}, x_{i2}, \ldots, x_{iF})$, $i = 1, \ldots, n$. We can minimize the mean squared error (MSE) as follows:

$$\min_{\beta_0, \beta_1, \ldots, \beta_F} \quad \frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_F x_{iF} - y_i)^2 \tag{1}$$

This is often called *the least squares problem* (first published in 1805!). The regression coefficients that minimize (1) are denoted $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_F$. The next section details how to mathematically determine the regression coefficients.

How do we determine the β's?

Let $\mathbf{y}^T = (y_1, y_2, \ldots, y_n)$, $\hat{\mathbf{y}}^T = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$, $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \ldots, \beta_F)$, and $\mathbf{X}$ be a full column rank $n \times K$ feature matrix, where each row represents a single observation (i.e., $(x_{i1}, x_{i2}, \ldots, x_{iK})$). Note that $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$ and recall that the Euclidean norm is defined as $||\mathbf{x}||_2 = (\sum_{i=1}^n |x_i|^2)^{1/2}$. We can rewrite (1) as:

$$\min_{\boldsymbol{\beta}} \quad \frac{1}{n}||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2.$$

This problem is an unconstrained convex minimization problem and has a closed-form solution. Let $\hat{\boldsymbol{\beta}}$ denote the (column) vector of optimal regression coefficients. We can expand the objective function to obtain:

$$||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 = (\mathbf{X}\boldsymbol{\beta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) = \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^T\mathbf{y}.$$

Therefore,

$$\min_{\boldsymbol{\beta}} \quad \frac{1}{n}||\mathbf{X}\boldsymbol{\beta} - \mathbf{y}||_2^2 \;\; = \;\; \min_{\boldsymbol{\beta}} \quad \frac{1}{n}(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - 2\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^T\mathbf{y}).$$

Taking the derivative with respect to $\boldsymbol{\beta}$ and equating to zero, we obtain the closed form solution for $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} 0 &= \frac{2}{n}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} - \frac{2}{n}\mathbf{X}^T\mathbf{y}, \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \end{aligned}$$

What are the assumptions of a linear regression model?

The basic assumptions for linear regression are as follows:

- **Weak exogeneity:** The feature values are known exactly and not subject to error. This must be assumed by the modeler.
- **Linearity:** The target can be written as a linear combination of features.
- **Constant variance or homoscedasticity:** The variance of the error terms is constant across all samples.
- **No autocorrelation:** The error terms are not correlated with each other. Note that this assumption is satisfied if the targets are assumed to be independent from each other.
- **Normality (or other distribution information):** The error terms can, but are not required, to be normally distributed. If they are, the regression coefficients are also normally distributed. This assumption impacts the statistical tests used for the regression coefficients and other distributional assumptions are possible.

How do we verify these assumptions?

It is imperative that we verify the model assumptions before using a linear regression model in an explanatory fashion. Residual (or error) plots can be used to test for linearity, homoscedasticity, autocorrelation, and normality. Residuals can also be adjusted or standardized to test for outliers. Note that a residual is defined as $r_i = \hat{y}_i - y_i$. Some common tests include:

- **Linearity, homoscedasticity, and autocorrelation:** We can check these assumptions using different residual plots:

  1. residuals (r_i) vs. predicted targets ($\hat{y}_i$)
  2. residuals (r_i) vs. case order ($i = 1, 2, ..., n$)

  In each plot, any patterns or an obvious non-random scatter indicates that there may be a violation and requires further investigation.

- **Normality:** We can check if the residuals are normally distributed using a histogram and a QQ-plot (a graphical method for determining if data is normally distributed). The histogram should appear symmetric and be centered at zero. The QQ-plot should have all values near the dashed line. Some variability is normal in the tails, but it should not be severe or pathogenic.

- **Outliers:** We can check for outliers using standardized residuals or metrics derived from the residuals. There are two commonly used metrics for assessing outliers:

  1. *Leverage:* measures features similarity for data points with (nearly) equal target values
  2. *Cook's distance:* measures the influence of each data point

  In each plot, large or outlying values may indicate an outlier. In some cases, significant outliers may need to be removed from the data to avoid biasing the results.

How do we interpret the model?

Some machine learning models, including linear regression, allow us to carefully explain the relationship between each feature and the target. To do this, we need to interpret the regression coefficients, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_F$. Consider the wine quality example. We will use the same data set that Ashenfelter used to train his original model, which includes 25 observations from 1952 to 1978 (1956 is missing) and four features (age, winter rain, harvest rain, average growing season temperature (AGST)). Using the statsmodel package in Python, we obtain the following model output:

OLS Regression Results

Dep. Variable:	Price	R-squared:	0.829
Model:	OLS	Adj. R-squared:	0.794
Method:	Least Squares	F-statistic:	24.17
Date:	Tue, 19 Feb 2019	Prob (F-statistic):	2.04e-07
Time:	20:36:53	Log-Likelihood:	-2.1622
No. Observations:	25	AIC:	14.32
Df Residuals:	20	BIC:	20.42
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
WinterRain	0.0011	0.001	2.120	0.047	1.73e-05	0.002
AGST	0.6072	0.099	6.152	0.000	0.401	0.813
HarvestRain	-0.0040	0.001	-4.652	0.000	-0.006	-0.002
Age	0.0239	0.008	2.956	0.008	0.007	0.041
const	-3.4300	1.766	-1.942	0.066	-7.114	0.254

Omnibus:	1.814	Durbin-Watson:	2.797
Prob(Omnibus):	0.404	Jarque-Bera (JB):	1.041
Skew:	0.034	Prob(JB):	0.594
Kurtosis:	2.003	Cond. No.	1.91e+04

- The *"coef"* column displays the regression coefficients $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_F)$
- The *"$P > |t|$"* column displays the p-value corresponding to each regression coefficient and the following two columns display the 95% confidence interval. We typically use a significance level of 5% when interpreting the coefficients, meaning that if the p-value is below 0.05, then the regression coefficient is *statistically significant*. If the p-value is above 0.05, then the 95% confidence interval includes zero, meaning that the regression coefficient is not statistically different from zero. We can only interpret features with regression coefficients that are statistically significant. However, statistically insignificant features may still contribute to the model's predictive accuracy!

In our wine example, we find that all four features are statistically significant. Consider the Age feature, which has a coefficient of 0.0239. Recall that the target is $log(Price)$ (measured in hundreds of US dollars). When we predict the log of a target variable, the interpretation of the coefficient changes: all other things being equal (ceteris paribus), for every one year increase in the age of a wine, the price of wine increases by approximately 2.39% (or exactly 2.42%). We arrive at this conclusion because a one unit change in age, increases the $log(Price)$ by 0.0239 or the price by $e^{0.0239} = 1.0242$ (and $e^{\beta} \approx 1 + \beta$ when β is small). If the target was $Price$ (without the logarithm), then a coefficient of 0.0239 implies that for every one year increase in the age of a wine, the price of wine increases by \$0.24.