Introduction

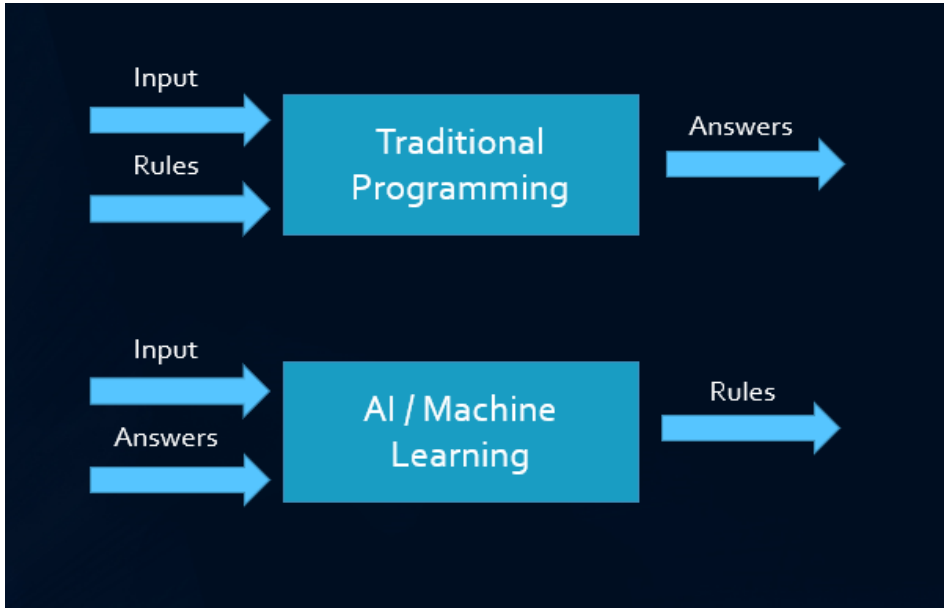
About this course

- *Introduction* to machine learning.
- 10 lectures covering a broad range of topics.
- It's a mathematical course (supplemented by practical exercises).
- The necessary math will be introduced as we go.

  For more, see Deisenroth et al., *Mathematics for Machine Learning* (free online).
- The focus is on introducing key concepts and developing intuitions.

What is machine learning?

"Machine learning (ML) is the study of computer algorithms that improve automatically through experience." (Wikipedia)



From the internet. Original source unknown.

What is machine learning?

"The goal of machine learning is to develop methods that can automatically detect patterns in data, and then to use the uncovered patterns to predict future data or other outcomes of interest." (Murphy, *Machine Learning*)

"Machine learning is thus closely related to the fields of statistics and data mining, but differs slightly in terms of emphasis [...]"

Machine learning vs. statistics

Both statistics and machine learning aim to "detect patterns in data" by building predictive models.

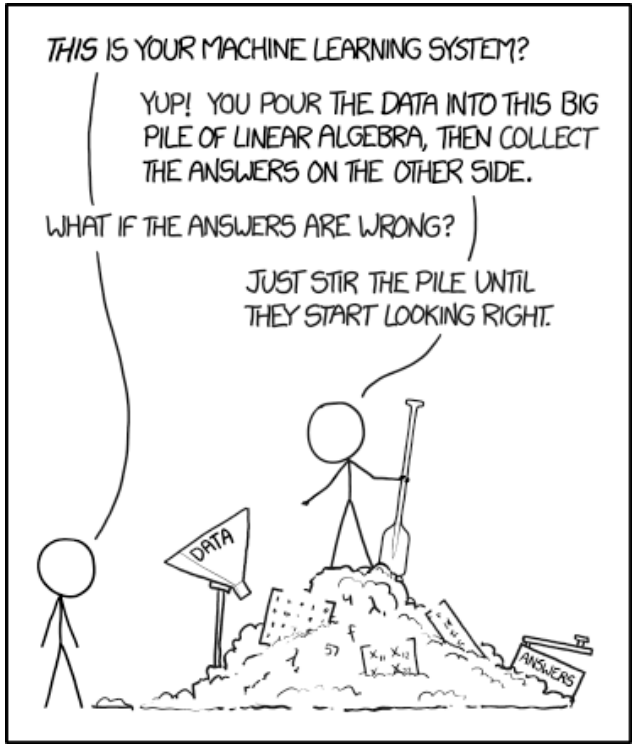
Statistics: use this as a tool to learn something about the world (*statistical inference*). Focus on simple, interpretable models. Develop theoretical analysis, work out statistical guarantees under some assumptions.

Machine learning: use this as a tool to actually make useful predictions. Focus on complicated, competitive models. Use large datasets. Be pragmatic. Give up on inference.

Breiman (2001) Statistical Modeling: The Two Cultures.

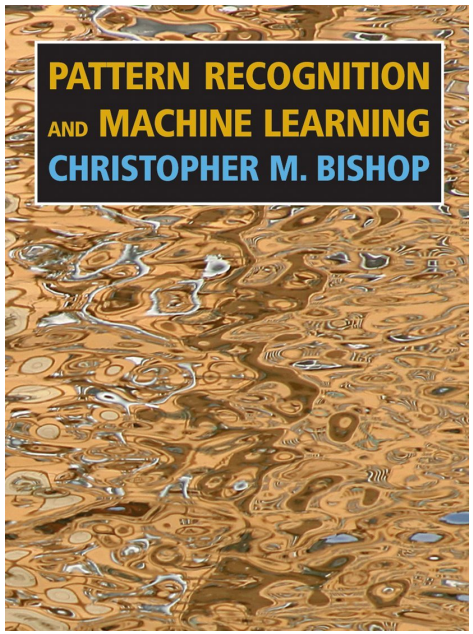
Machine learning vs. statistics: spectrum

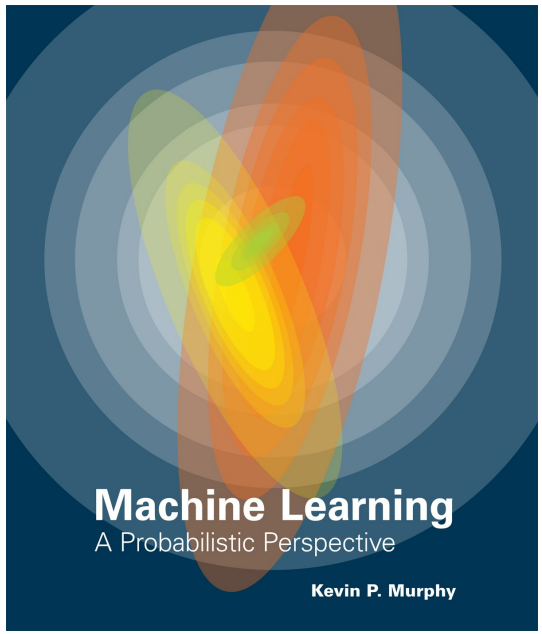
In practice, there is no boundary — it is a spectrum. It goes all the way from a one-sample t-test to GPT-3.

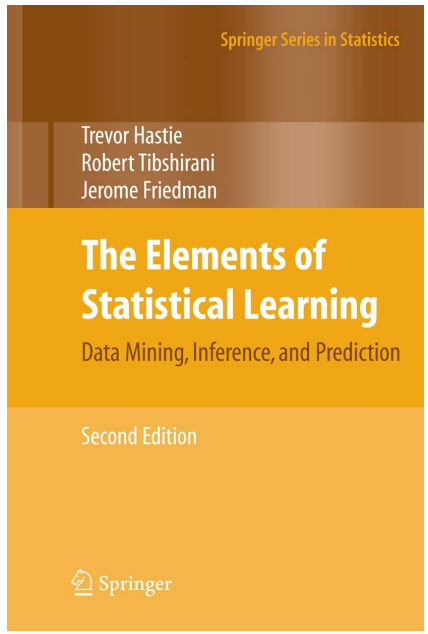


https://xkcd.com/1838/

“Statistical learning”







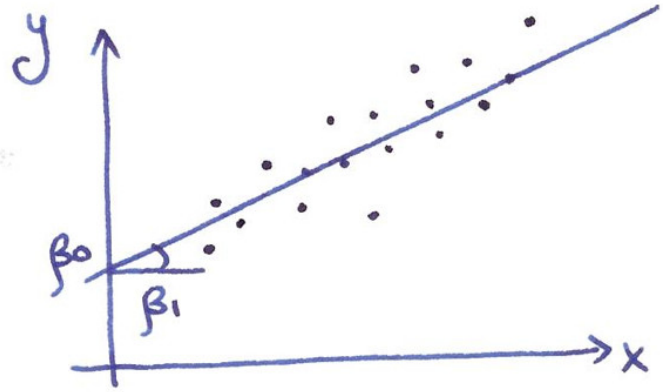
Types of machine learning problems

1. Supervised learning
   Example: distinguish photos of cats from photos of dogs.
2. Unsupervised learning
   Example: figure out that cat and dog photos show different animals.
3. Reinforcement learning
   Example: play Go.

Yann LeCun: "Most of human and animal learning is unsupervised learning. If intelligence was a cake, unsupervised learning would be the cake, supervised learning would be the icing on the cake, and reinforcement learning would be the cherry on the cake."

Simple linear regression

Simple linear regression



Supervised learning problem. Regression (not classification) problem.

Training data: $\{(x_i, y_i)\}_{i=1}^{n}$.

Model: $\hat{y} = f(x) = \beta_0 + \beta_1 x$.

Two coefficients: *intercept* and *slope*. We want to *fit* the model to the data.

Loss function

To fit the model means to find β_0 and β_1 so that $f(x_i) \approx y_i$.

Loss function (aka cost function):

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

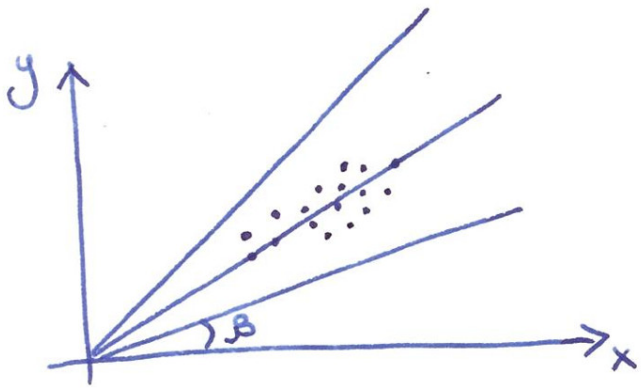
Mean squared error (MSE). Why MSE?

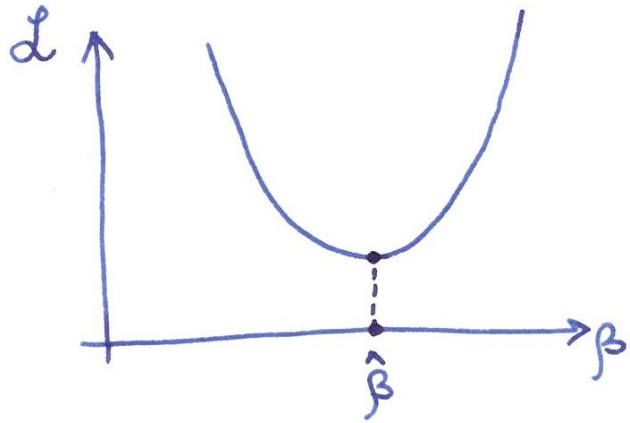
Ordinary least squares (OLS).

Baby linear regression

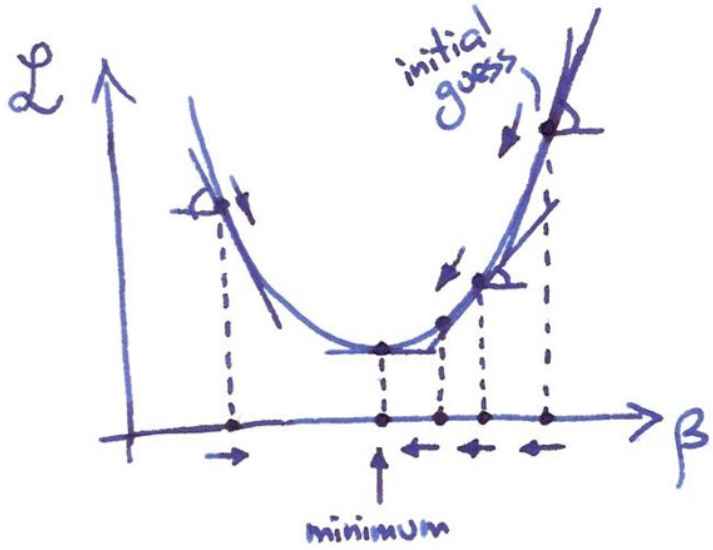
Consider slope-only model: $f(x) = \beta x$.

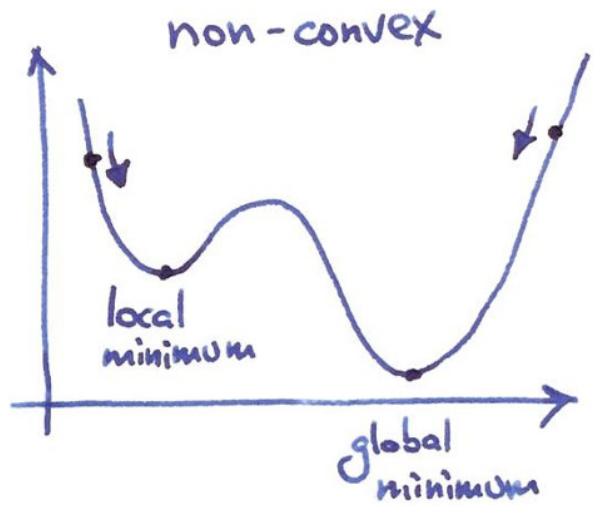
The loss: $\mathcal{L}(\beta) = \frac{1}{n}\sum_i (y_i - \beta x_i)^2$.





Baby gradient descent





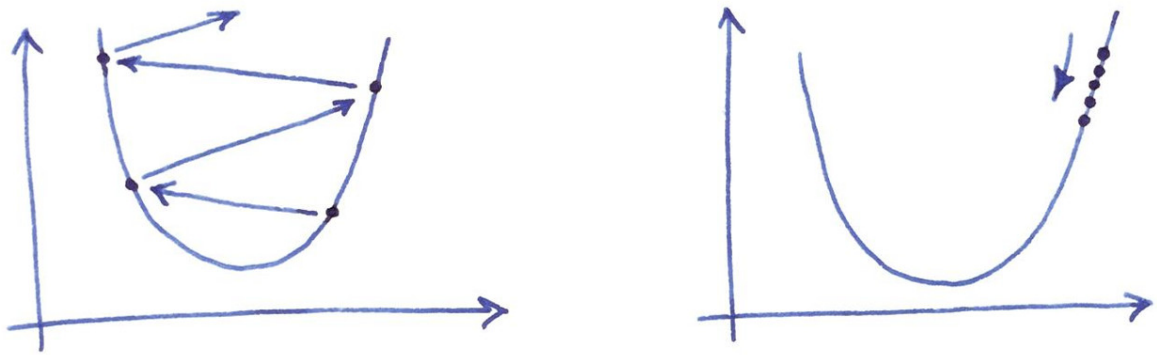
Update rule:

$$\beta \leftarrow \beta - \eta \frac{d\mathcal{L}(\beta)}{d\beta}.$$

Here η is the learning rate.

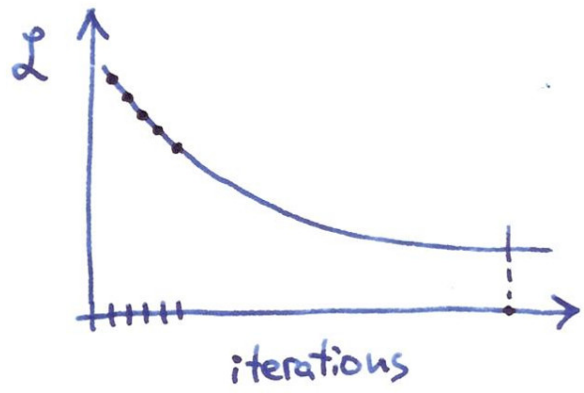
Learning rate

$$\beta \leftarrow \beta - \eta \frac{d\mathcal{L}(\beta)}{d\beta}$$



Too large η — divergence. Too small η — slow convergence.

Stopping criterion



$$\beta \leftarrow \beta - \eta \frac{d\mathcal{L}(\beta)}{d\beta}$$

Baby gradient descent cont.

We need to compute the derivative of the loss:

$$\mathcal{L}(\beta) = \frac{1}{n}\sum_i (y_i - \beta x_i)^2.$$

We get:

$$\mathcal{L}'(\beta) = \frac{1}{n}\sum_i 2(y_i - \beta x_i)(-x_i) = -\frac{2}{n}\sum_i x_i(y_i - \beta x_i).$$

Baby analytical solution

$$\mathcal{L}'(\beta) = -\frac{2}{n}\sum_i x_i(y_i - \beta x_i).$$

At the minimum:

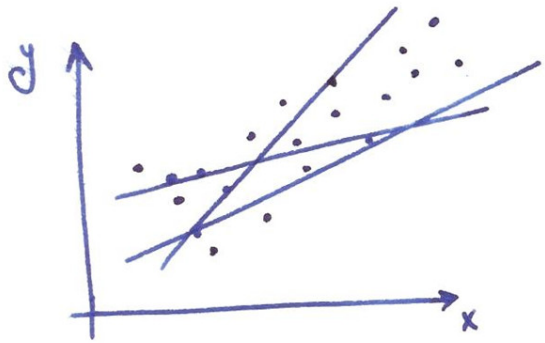
$$\sum_i x_i y_i - \hat{\beta}\sum_i x_i^2 = 0.$$

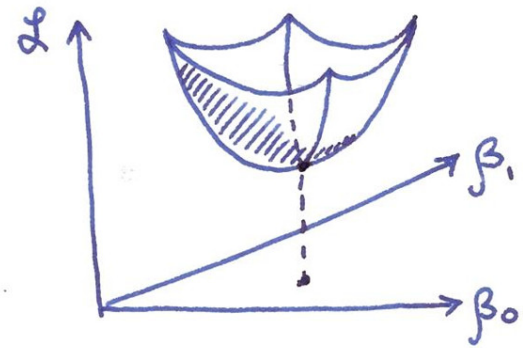
We obtain:

$$\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}.$$

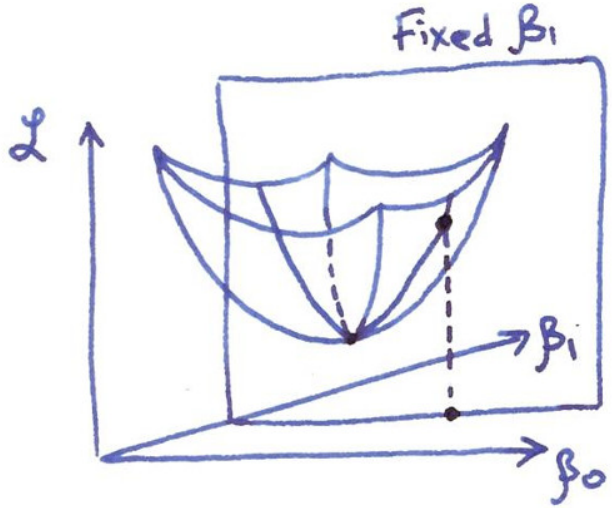
Back to simple linear regression

$$\mathcal{L}(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$





Introducing *partial derivatives*



$$\beta_0 \leftarrow \beta_0 - \eta \frac{\partial \mathcal{L}}{\partial \beta_0}$$

$$\beta_1 \leftarrow \beta_1 - \eta \frac{\partial \mathcal{L}}{\partial \beta_1}$$

Introducing *gradient*

Update rules for each parameter:

$$\beta_0 \leftarrow \beta_0 - \eta \frac{\partial \mathcal{L}}{\partial \beta_0}$$

$$\beta_1 \leftarrow \beta_1 - \eta \frac{\partial \mathcal{L}}{\partial \beta_1}$$

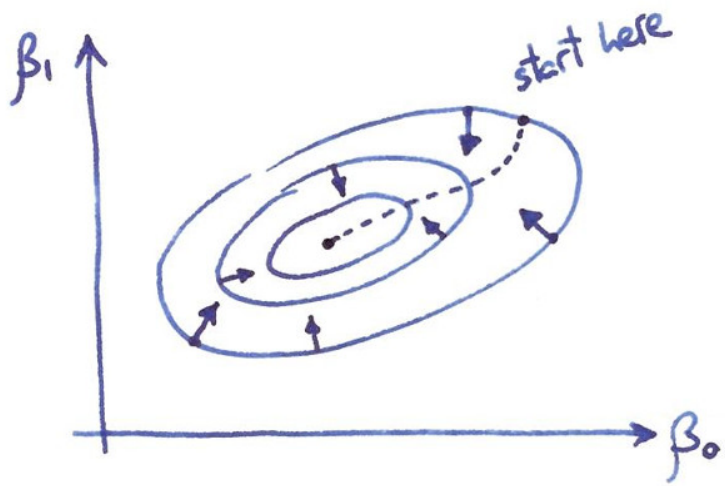
In vector form:

$$\vec{\beta} \leftarrow \vec{\beta} - \eta \nabla \mathcal{L}.$$

Gradient descent

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

$$\vec{\beta} \leftarrow \vec{\beta} - \eta \nabla \mathcal{L}$$



Computing the gradient

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

We need partial derivatives:

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = -\frac{2}{n}\sum(y_i - \beta_0 - \beta_1 x_i)$$
$$\frac{\partial \mathcal{L}}{\partial \beta_1} = -\frac{2}{n}\sum(y_i - \beta_0 - \beta_1 x_i)x_i$$

Exercise: derive the analytical solution for $\hat{\beta}_0$ and $\hat{\beta}_1$.