Introduction to Machine Learning

CMPUT 328

Nilanjan Ray

Computing Science, University of Alberta, Canada

Material source: “Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems,” by Géron, Aurélien.

What is machine learning?

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

-Arthur Samuel, 1959

A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

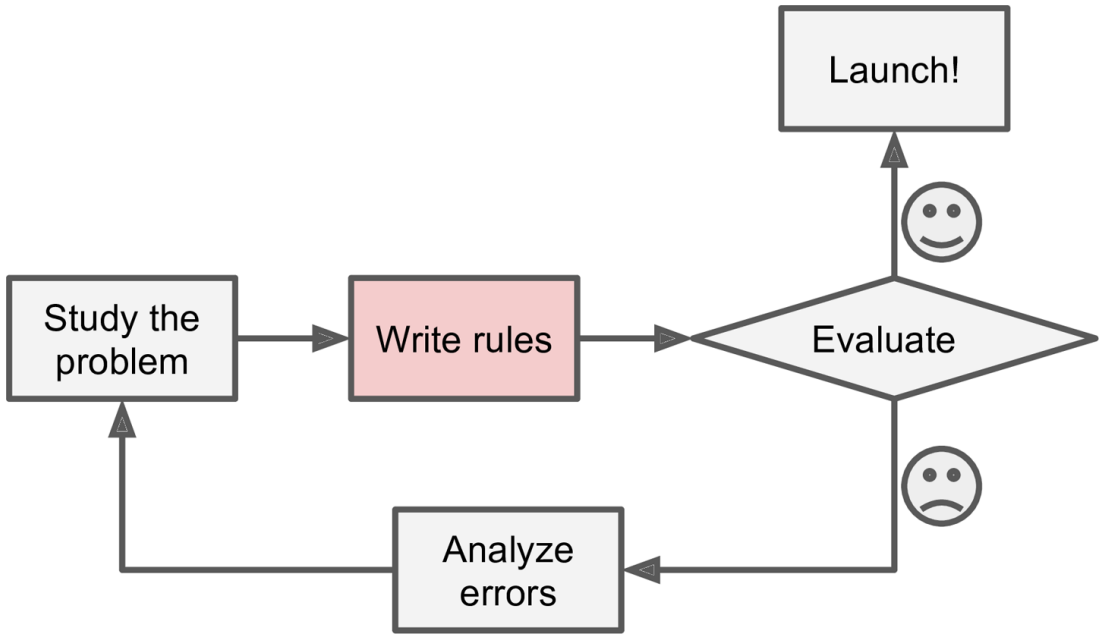
-Tom Mitchell, 1997

Why do we need machine learning?

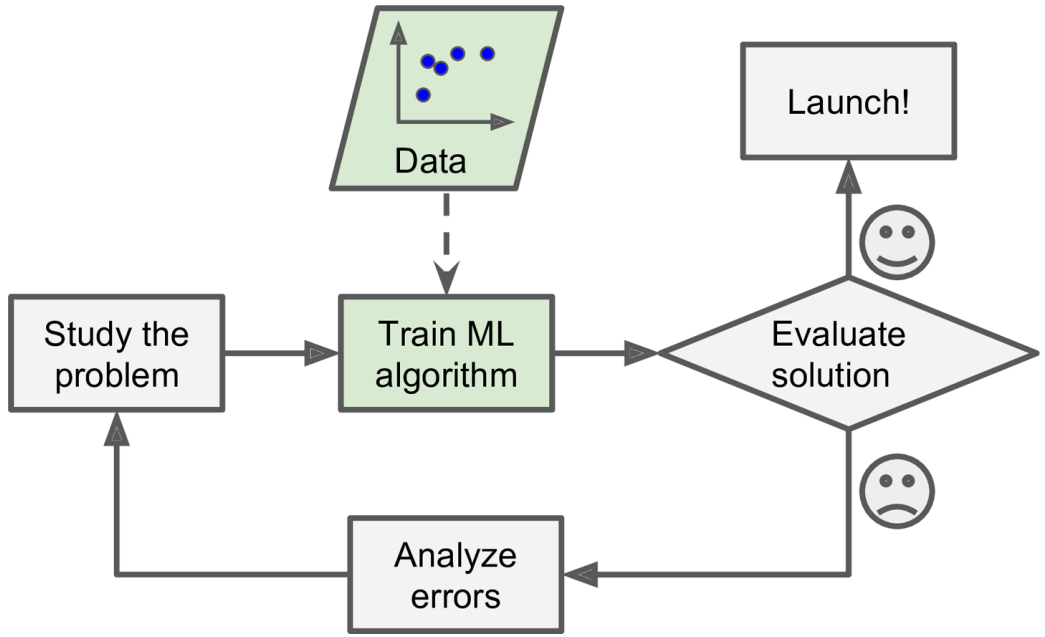
- Consider the spam filter example in your email.

Traditional vs. machine learning approach

Traditional view



ML view



Summary: why we need ML

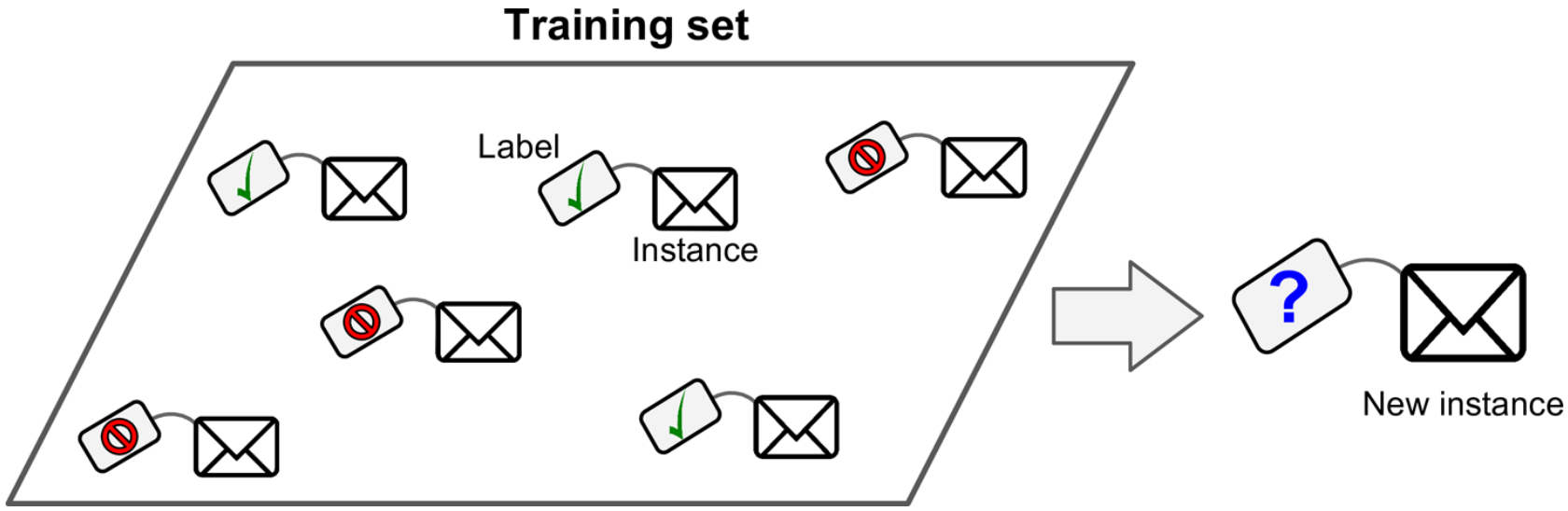
- Problems for which existing solutions require a lot of hand-tuning or long lists of rules: one Machine Learning algorithm can often simplify code and perform better.
- Complex problems for which there is no good solution at all using a traditional approach: the best Machine Learning techniques can find a solution.
- Fluctuating environments: a Machine Learning system can adapt to new data.
- Getting insights about complex problems and large amounts of data.

Types of ML

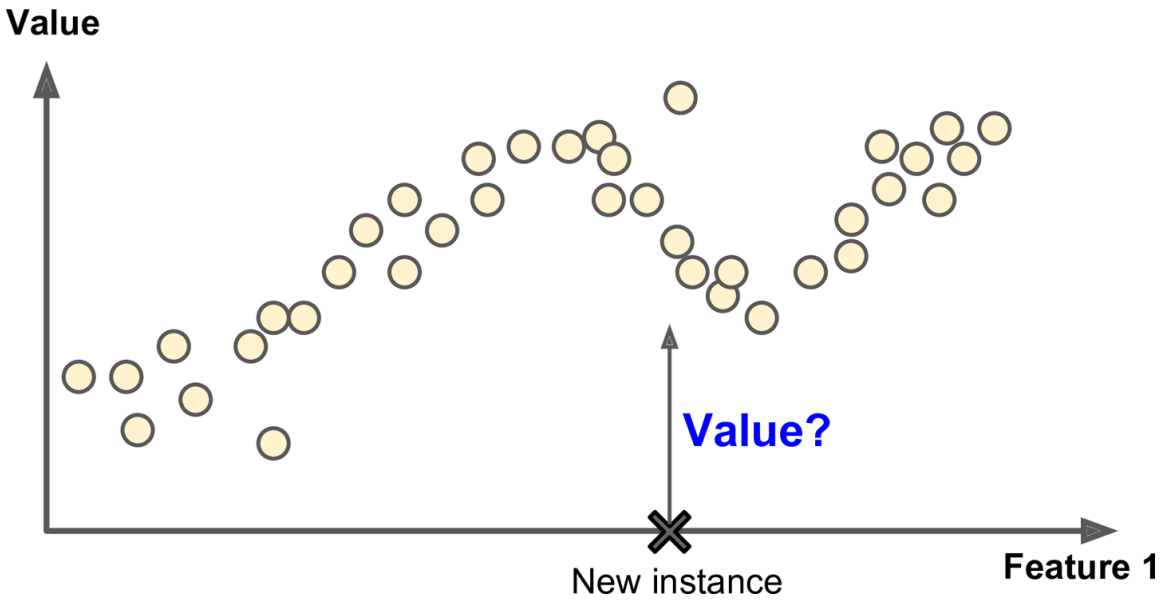
- Whether or not they are trained with human supervision (supervised, unsupervised, semisupervised, and Reinforcement Learning)
- Whether or not they can learn incrementally on the fly (online versus batch learning)
- Whether they work by simply comparing new data points to known data points, or instead detect patterns in the training data and build a predictive model, much like scientists do (instance-based versus model-based learning)
- These criteria are not exclusive; you can combine them in any way you like. For example, a state-of-the-art spam filter may learn on the fly using a deep neural network model trained using examples of spam and ham; this makes it an online, model-based, supervised learning system.

Supervised learning

Classification:
The spam filter is a good example of this: it is trained with many example emails along with their *class* (spam or ham), and it must learn how to classify new emails.

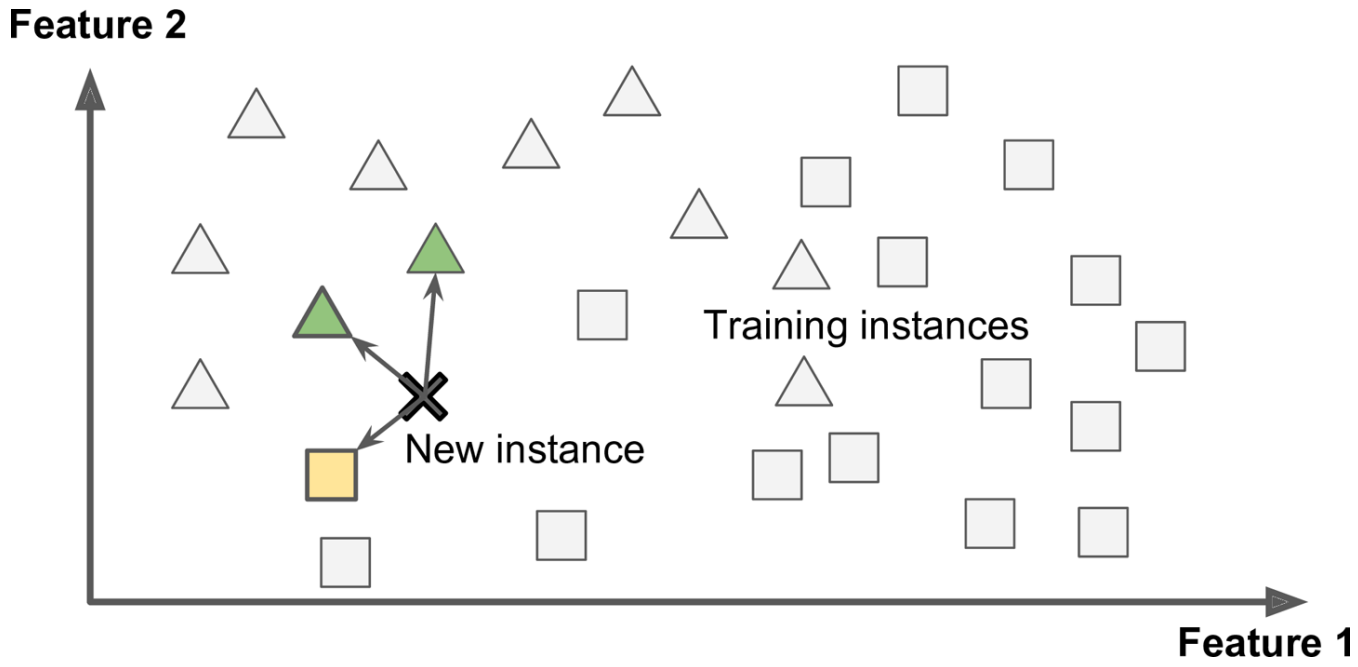


Regression: Another typical task is to predict a *target* numeric value, such as the price of a car, given a set of *features* (mileage, age, brand, etc.) called *predictors*. To train the system, you need to give it many examples of cars, including both their predictors and their labels (i.e., their prices).



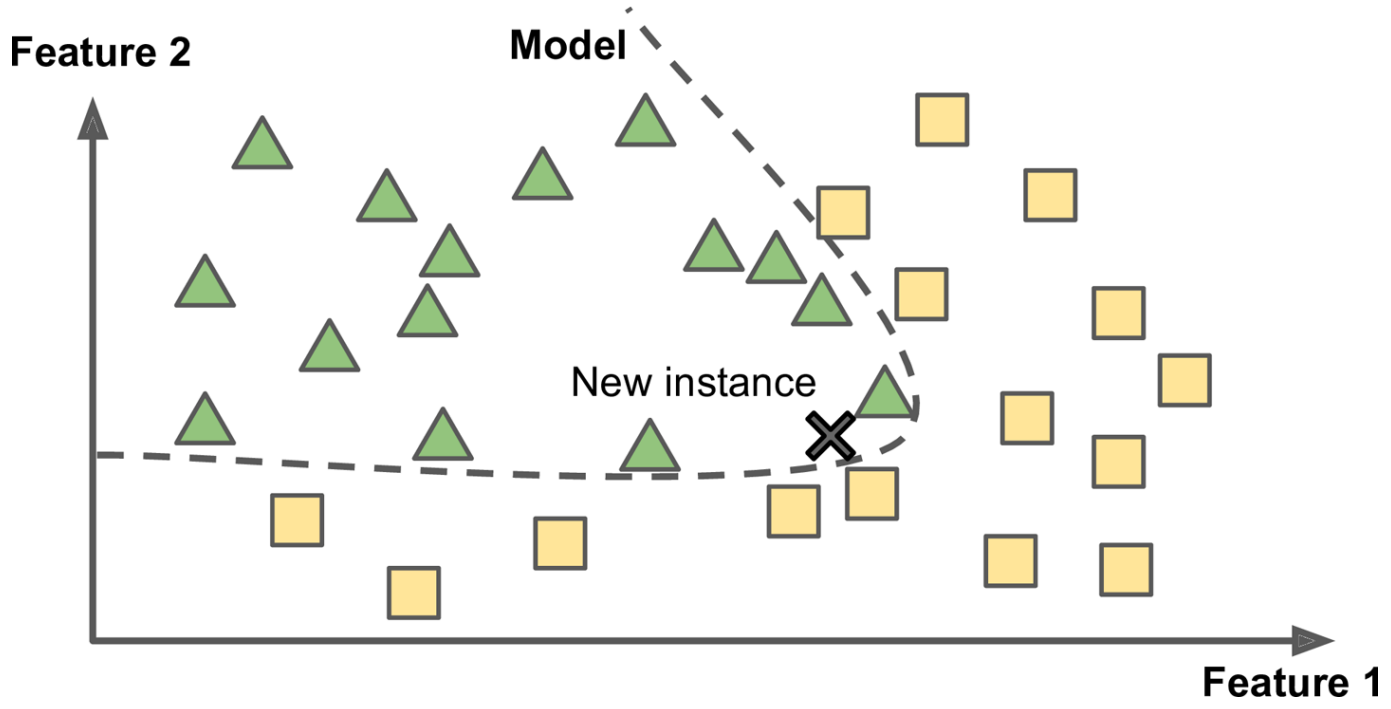
Instance-based supervised learning

- Remember all training examples
- When a test email comes, compare it with its “neighbors” from the training examples and classify accordingly
- Requires a measure of similarity
- Example: k-nearest neighbor (knn) method



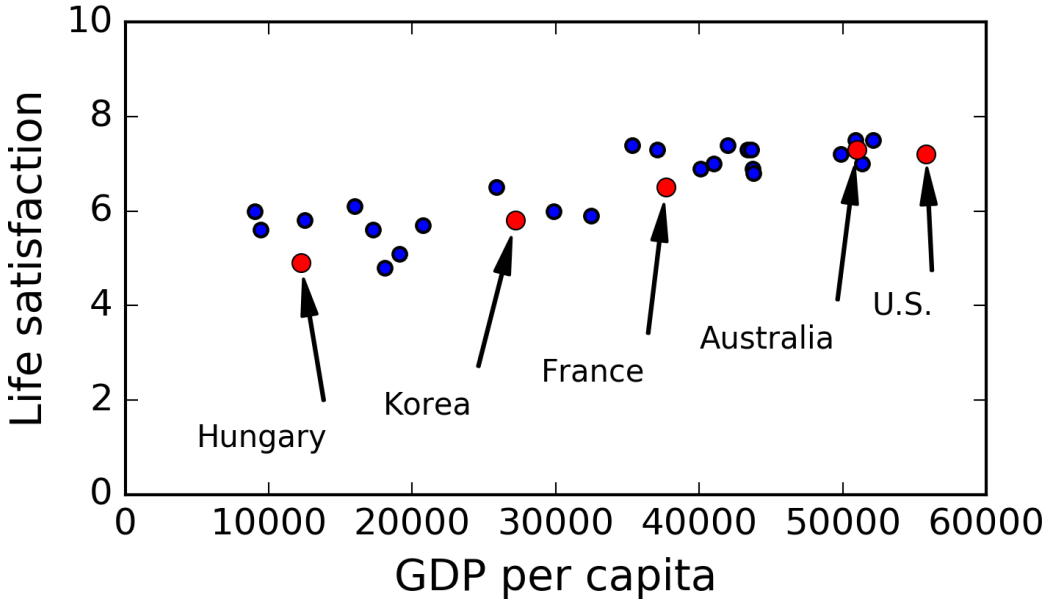
Model-based supervised machine learning

- From all the training examples, build a model for the learner
- When a test example comes, apply the model
- Don’t need to remember all training examples, after training
- Examples: neural net, support vector machine, linear regression, etc.

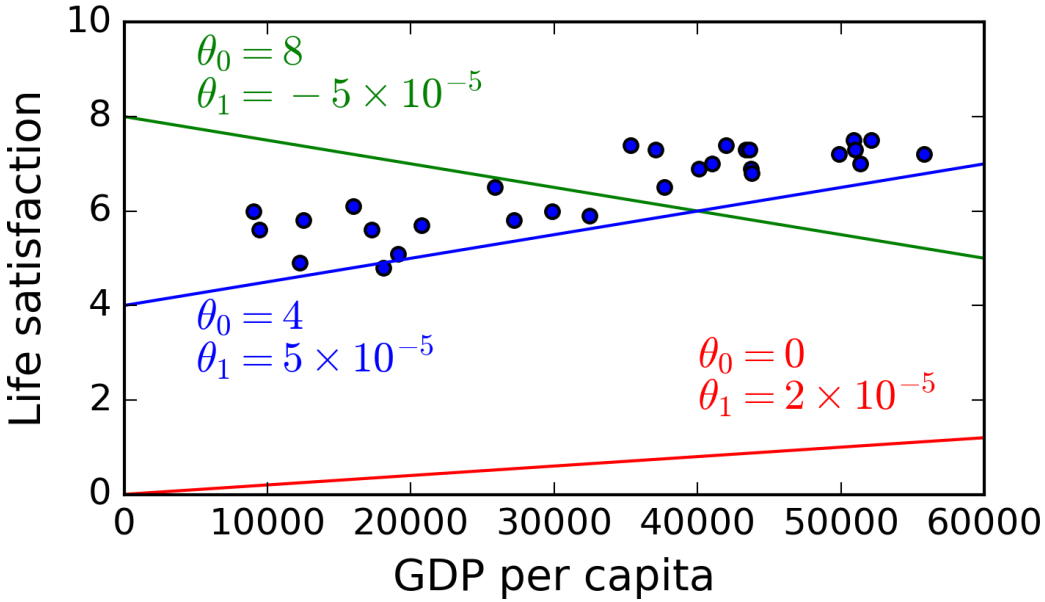


Linear-model based supervised learning

Do we see any trend?



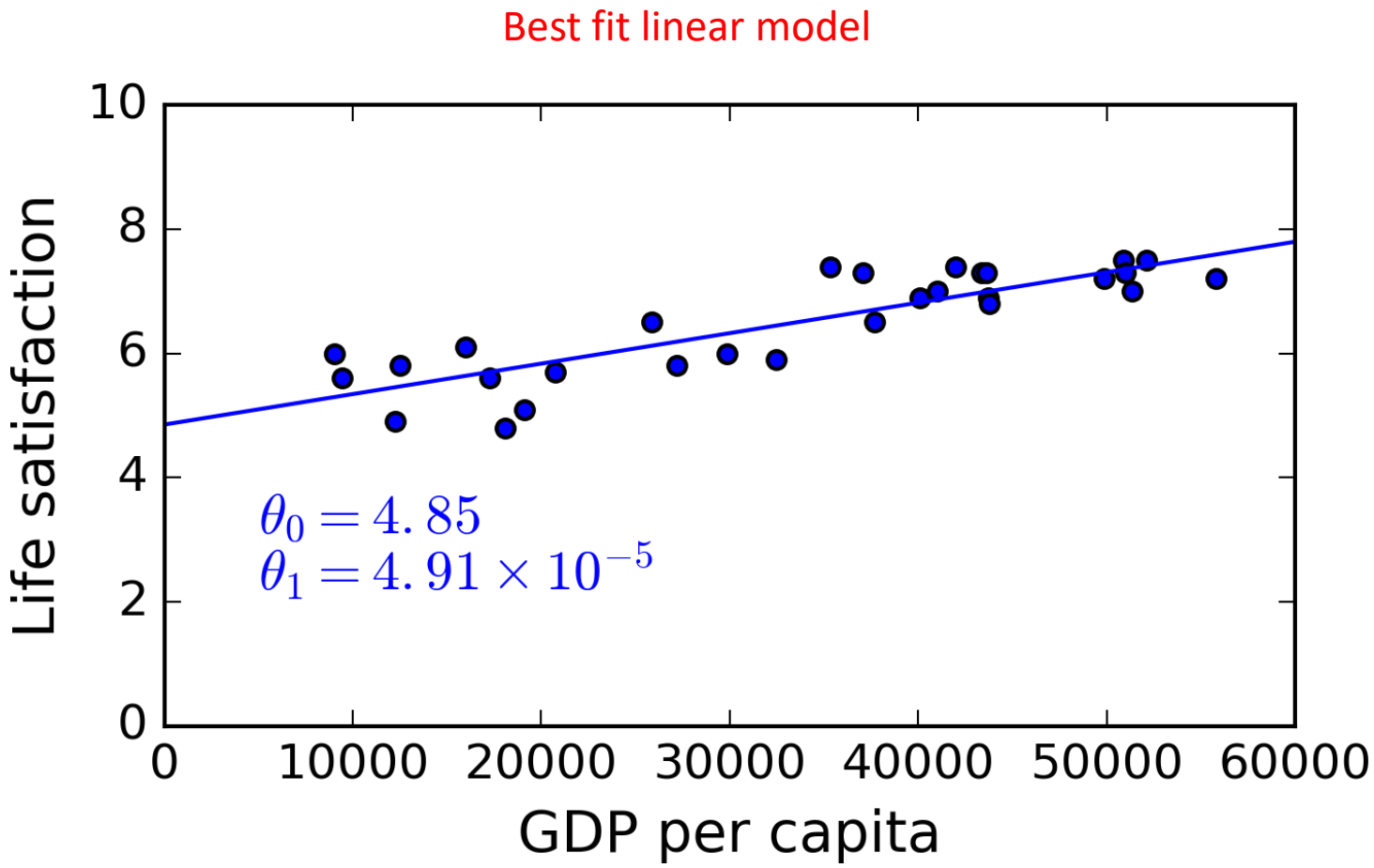
A few possible linear models



Linear model: $\text{life\_satisfaction} = \theta_0 + \theta_1 \times \text{GDP\_per\_capita}$

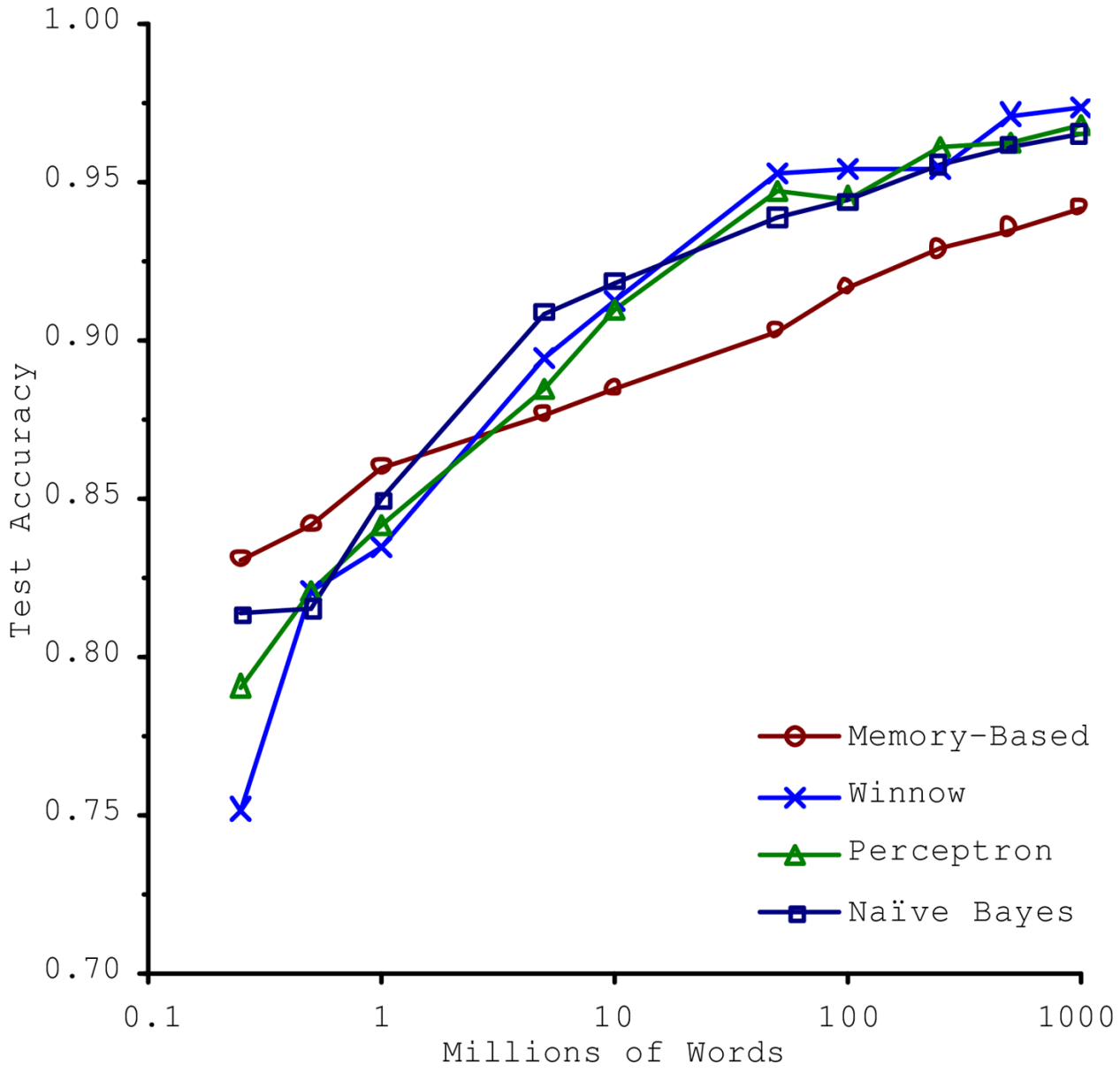
Parameters of the model: θ_0, θ_1

Linear-model based supervised learning



Challenge 1

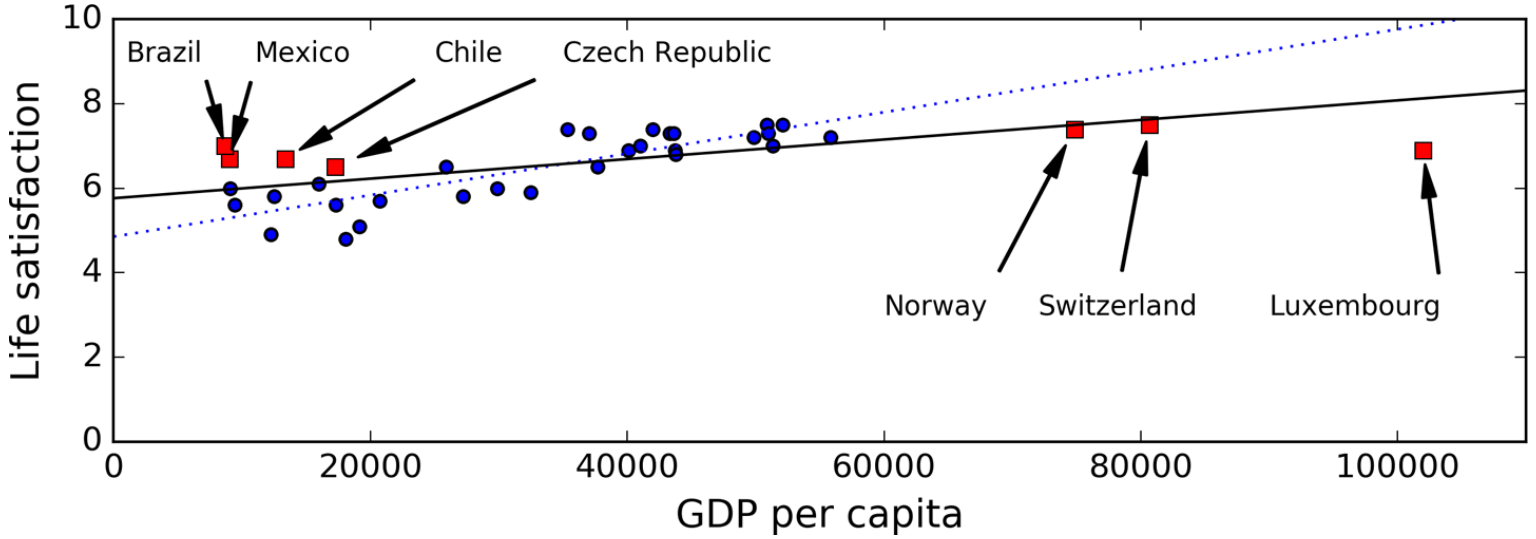
- Insufficiency of labeled training data



The importance of data versus algorithms: by Peter Norvig

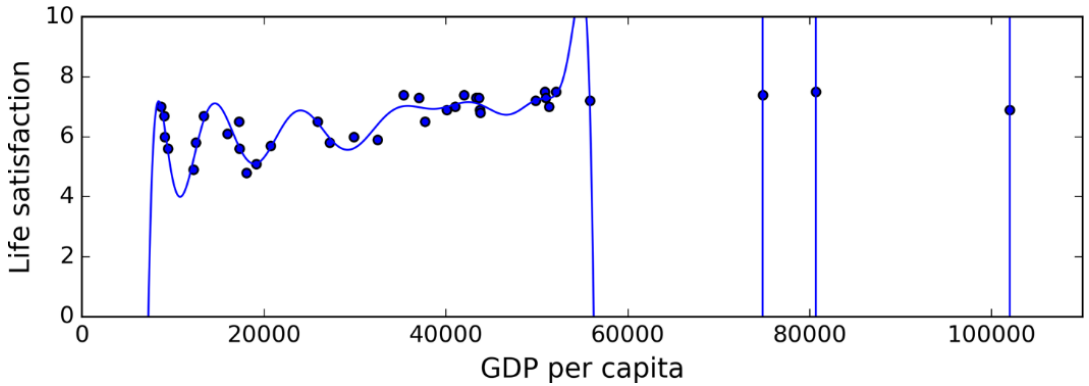
Challenge 2

- Non-representative training data

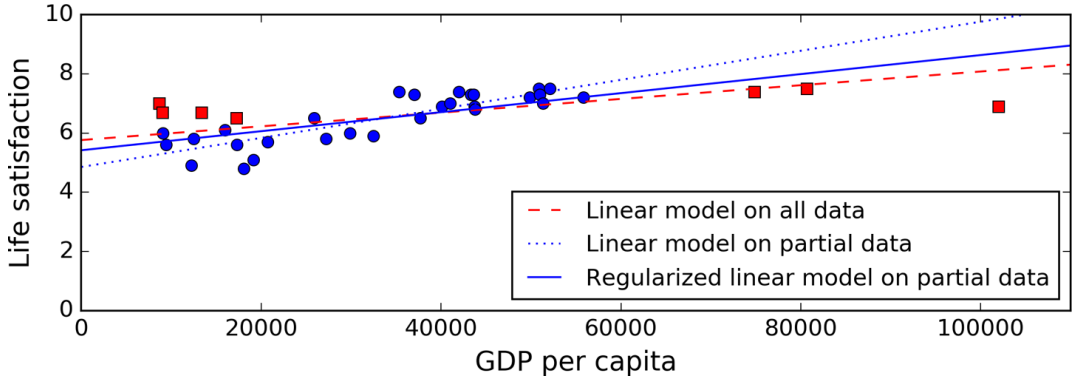


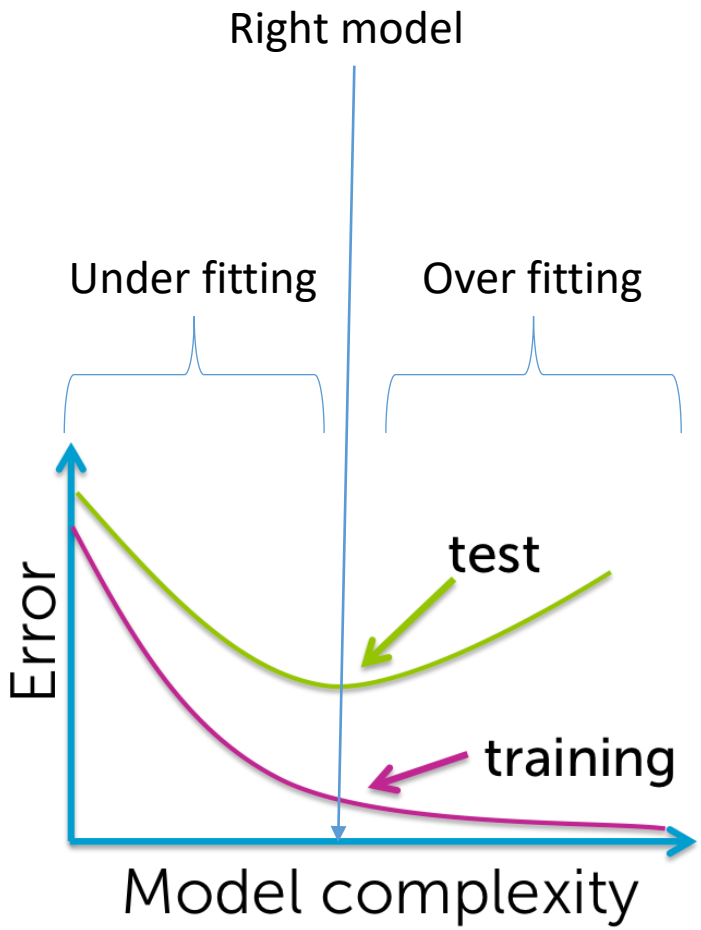
Challenge 3

Fitting a high degree polynomial: typical overfitting



Regularization reduces risk of overfitting





What is regularization?

Testing and validating model

- No free lunch theorem: David Wolpert demonstrated that if you make absolutely no assumption about the data, then there is no reason to prefer one model over any other.
- In practice, you have to try a model to your data to see how well it performs
- Divide data into training, validation and test. We will see how to use these in the practical session.