**LECTURE 1:**

* **A learning problem** can be described as a triple (P, T, E), P is performance, T is task, E is experience

**Example:** Spam filtering for emails

+ T: filter/predict all emails that are spam.

+ P: the accuracy of prediction, that is the percentage of emails that are correctly classified into normal / spam.

+ E: set of old emails, each with a label of spam/normal.

* **Two basic learning problems:**
* **Supervised learning:**

Classification: Multiclass, Multilabel

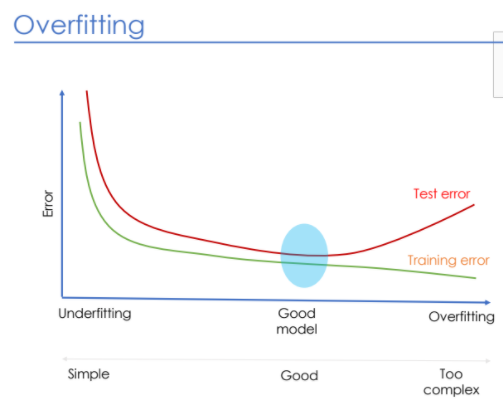
Regression

* **Unsupervised learning:**

Retrieval

Clustering

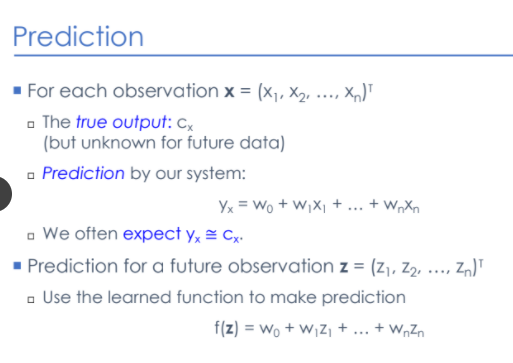
* **Design a learning system**
* Model
* Algorithm
* **Overfitting:** It is more accurate in fitting known data, but less accurate in predicting unseen data. Reasons:
* The function/model is too complex or have too much parameters.
* Noises or errors are present in the training data.
* The training size is too small, not characterizing the whole space.



* **Regularization:** Reduce Overfitting
* **Tiền xử lý:**
* **Thu thập dữ liệu:** crawling, logging, scraping
* **Xử lý dữ liệu:** lọc nhiễu, số hóa, lưu trữ

**LECTURE 2:**

* **Regression problem:** predict f(xi)
* **Linear model:**

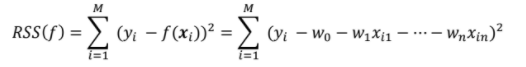


* **Loss function:**
* **Error / Loss:**



* **Empirical loss / Expected loss ~ risk** (dùng khi có xác suất p(x, y))





* **Generalization error:**