Machine Learning
EDS
Formalising the Prediction Problem, Risk and Ideal Predictors

Janneke van Brummelen

Vrije Universiteit Amsterdam

Period 3.4
2023/2024

Table of contents

Table of Contents

Data

In **supervised learning**, we assume to have access to a **sample of n examples**

$$D_n = (X_i, Y_i)_{1 \leq i \leq n}$$

where for all $i \in \{1, \ldots, n\}$

- $X_i \in \mathcal{X}$ is an *explanatory variable* or *feature*
- $Y_i \in \mathcal{Y}$ is a *variable of interest*, or *label*

Assumption throughout the course:

- $(X, Y), (X_1, Y_1), (X_2, Y_2), \ldots$ **independent and identically distributed** random variables with **common distribution** P
- Technical assumption: $\mathcal{X}$, $\mathcal{Y}$ measurable spaces

Data: remark

Often, the X_i's are actually a collection of several explanatory variables concatenated in a vector of $\mathbb{R}^p$

Example:

- $Y_i = \text{income} \in \mathbb{R}$
  $X_i = (\text{age}, \text{age}^2) \in \mathbb{R}^2$

- $Y_i = \text{Presence of a pedestrian on a colour photo} \in \{0, 1\}$
  $X_i = \text{Collection of RGB intensity pixels} \in [0, 1]^{3 \times \text{number of pixels}}$

What most problems reduce to

- The nature of the predicted variable Y characterises the type of learning problem
- Most supervised learning problems boil down to either:
  1. Predicting Y; a category/label/discrete outcome
  2. Predicting Y; a continuous value or a vector of continuous values

Discrete outcomes : Classification

- **Binary prediction** problems
  - "Cat" or "Dog" : 1 or 0
  - "Pedestrian" or "No pedestrian": 1 or 0
  - "Stocks will go up" or "Stocks will go down" : 1 or 0
  - "Cancerous cell" or "Not Cancerous cell" : 1 or 0
- **Multiclass problems** can be broken down into binary problems:
  - Multi-class problem : "Cat", "Dog", or "Elephant" 0, 1, 2
  - One-vs-All problem:
    - Sub-problem 1: "Cat" vs "No Cat" 1 or 0
    - Sub-problem 2: "Dog" vs "No Dog" 1 or 0
    - Sub-problem 3: "Elephant" vs "No Elephant" 1 or 0
  - One-vs-One problem:
    - Sub-problem 1: "Cat" vs "Dog" 1 or 0
    - Sub-problem 2: "Cat" vs "Elephant" 1 or 0
    - Sub-problem 3: "Dog" vs "Elephant" 1 or 0

Continuous outcomes : Regression

- Predicting a **continuous variable**
  - Steering wheel angle $\theta \in [-\pi, \pi]$
  - Car Speed $v \in \mathbb{R}$
  - Stock Price $P_t \in [0, \infty)$
  - Temperature $T \in [-273.15\text{C}^\circ, \infty)$

Classification and Regression

Classification and regression

Classification problem: $\mathcal{Y}$ is a finite set
Without loss of generality: $\mathcal{Y} = \{0, 1, 2, \ldots, k\}$

Binary classification problem: $\mathcal{Y}$ contains only two elements
Without loss of generality: $\mathcal{Y} = \{0, 1\}$

Regression problem: $\mathcal{Y}$ infinite set
Without loss of generality: $\mathcal{Y} = \mathbb{R}$ (or $\mathbb{R}^p$)

Vocabulary confusion pitfall

"Linear Regression" in econometrics ≠ "Regression" in ML

- **Linear Regression in econometrics**:
  Fitting a linear model $Y = a_1 X_1 + \ldots + a_p X_p + \varepsilon$
- **Regression in ML**:
  Any learning problem where predicted variable Y is continuous
- Ex: Learning to predict stock prices (Y, continuous variable) given any features (X) is a regression problem, *irrespective of the algo/model* (neural network, SVM, linear...)

Binary classification and Regression

- Binary classification and Regression are the cornerstones of supervised learning
- Many supervised learning problems can be reformulated into one or the other
- We will see them come back all the time

Data: Why random variables?

Why assume $(X_i, Y_i)_{1 \leq i \leq n}$ are random variables with law P?

- Randomness in X_i accounts for the random sampling of the examples from a given population
  - Ex: Images of handwritten numbers by a European / American
- Even given X_i, the variable Y_i can be random
  - Wrong labelling (A 3 mistakenly labelled as a 6)
  - Easily confused numbers (American 7 and European 1)
  - Handwriting that is hard to read

Note: The distribution P describes the (joint) distribution of the features and labels in a population and is typically *unknown*

Output of the learning problem

Predictor

A function $f : \mathcal{X} \longrightarrow \mathcal{Y}$ will be called a predictor and $\mathcal{F}$ will denote the set of all predictors

Solving the prediction problem amounts to find, using the observed sample $(X_i, Y_i)_{1 \leq i \leq n}$, a 'good' predictor f

'Good' typically refers to the generalisation ability of f:

- For a new observation X_{n+1}, one wishes that $f(X_{n+1})$ is 'close' to the (unobserved) true label Y_{n+1}

Other aspects can be important depending on the applications

- Computationally easy to find/learn/estimate f given a sample
- Computationally easy to evaluate $f(X_{n+1})$
- Interpretability of f
- Fairness of f

Remarks

Remarks

- A "predictor" is **any** function that maps a point x from the feature space to an output y.

  Nothing presumed about accuracy in the definition: a predictor could be very good or very poor.

- Predictors are also often called hypotheses. The set $\mathcal{F}$ is then said to be the set of all hypotheses.

Cost functions

To evaluate the quality of a given predictor, we need to measure "how far/close" its predictions are from the true labels

- To do so, we will consider a cost function, or loss, which can be any (measurable) function $c : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$
- Idea: the more similar $y, y' \in \mathcal{Y}$ are,
  the smaller $c(y, y')$ should be
  - The choice of c typically depends on the application
  - For simplicity, we will assume
    $\forall y, y' \in \mathcal{Y}, \quad c(y, y') \geq 0 \quad \text{and} \quad c(y, y) = 0$
- The goal is then to find a predictor $f \in \mathcal{F}$ such that $c(f(X_{n+1}), Y_{n+1})$ is small "on average"

Example of cost functions

- Regression: Y takes continuous values in $\mathbb{R}$
  $\forall y, y' \in \mathbb{R}$,
  - Quadratic cost: $c(y, y') = (y - y')^2$
  - Abs. value cost: $c(y, y') = |y - y'|$
  - L^p cost, $p > 1$: $c_p(y, y') = |y - y'|^p$
  - Truncated L^2 : $c_\alpha(y, y') = \min\{(y - y')^2, \alpha\}$, $\alpha > 0$
  - α-insensitive loss: $c_\alpha(y, y') = \max(0, |y - y'| - \alpha)$, $\alpha > 0$

- Binary classification: Y takes discrete values in $\{0, 1\}$
  $\forall y, y' \in \{0, 1\}$,
  - 0-1 cost: $c(y, y') = \mathbb{1}_{y \neq y'}$
  - Asymmetric cost:
    for some weights $w_0, w_1 \geq 0$, $w_0 + w_1 > 0$,

$$c_{w_0, w_1}(y, y') = w_{y'} \mathbb{1}_{y \neq y'}$$

Risk/generalisation error of a predictor

Definition: Risk/generalisation error of a predictor

Given an iid sample $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ with common distribution P, and a cost function c, the risk or generalisation error of a predictor $f \in \mathcal{F}$ is defined by

$$\mathcal{R}_P^c(f) = \mathbb{E}\Big[c(f(X), Y)\Big|D_n\Big],$$

where $(X, Y) \sim P$ is independent of D_n.

- The risk is a function from $\mathcal{F}$ to $\mathbb{R}$ depending on the chosen cost c and on the joint distribution P of features/labels
- In the above definition, $c(f(X), Y)$ is independent of D_n and the conditioning could be removed. Later, we will make f depend on the sample and conditioning on D_n will matter.

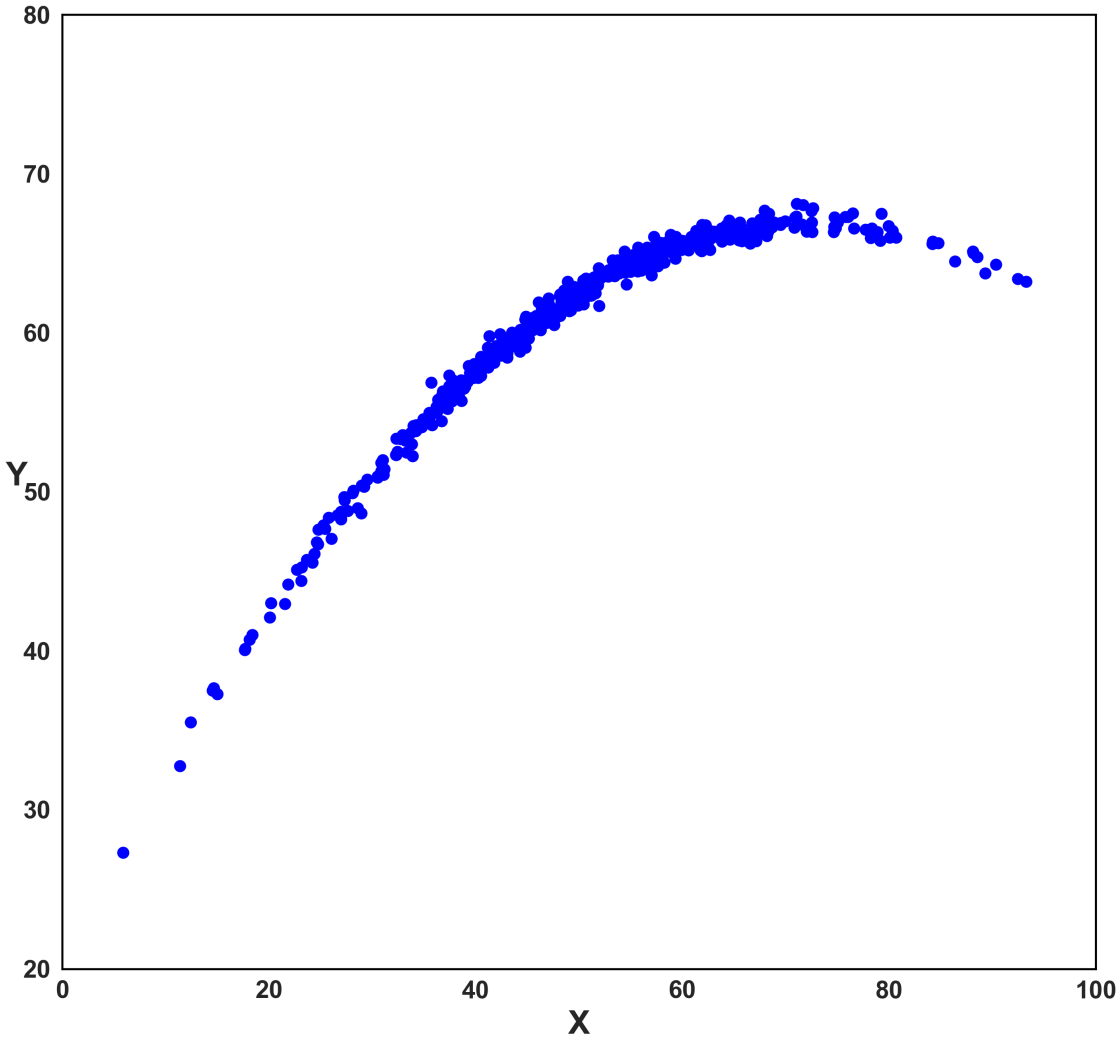
Formalisation of the learning problem

Learning problem

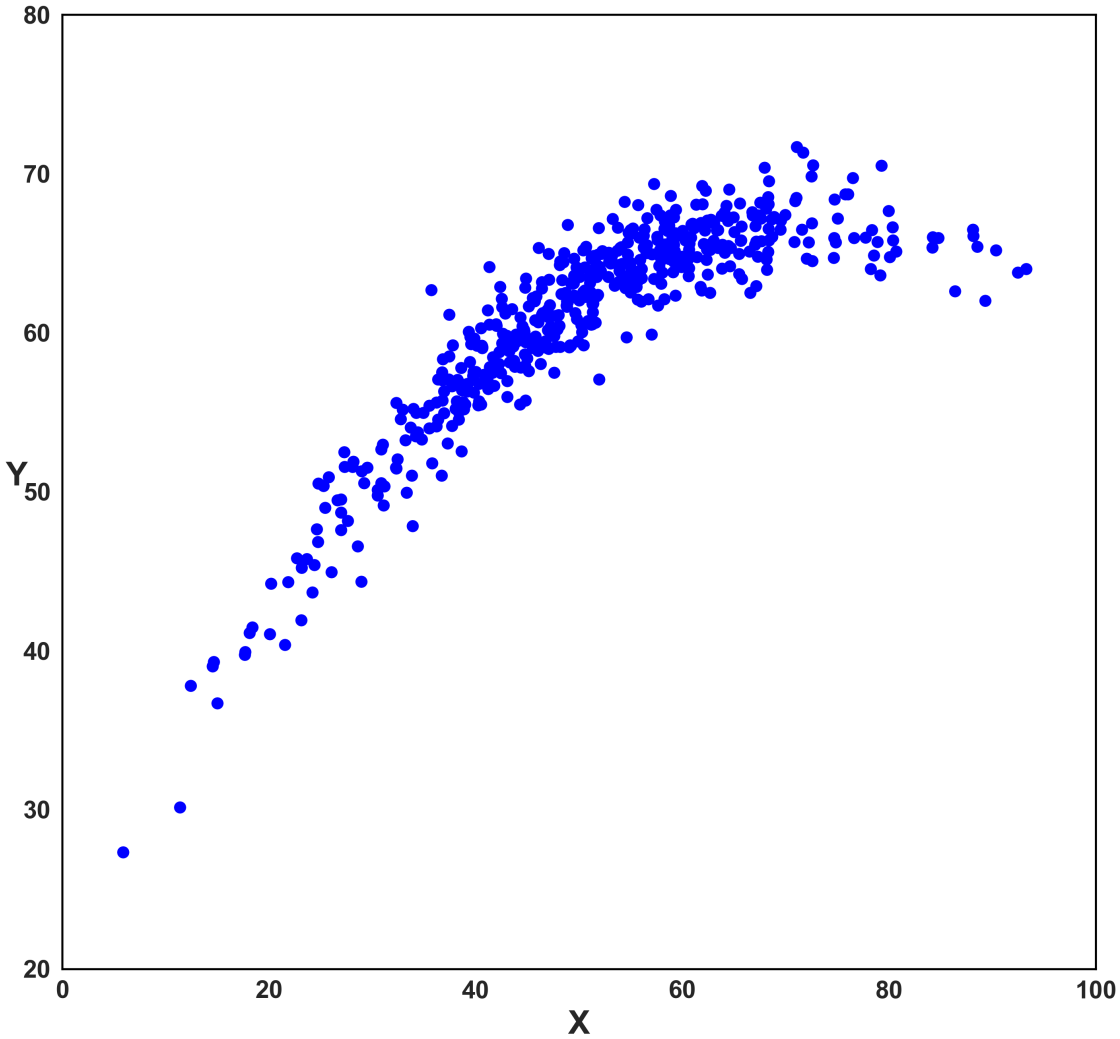
The learning problem consists of finding a predictor $f \in \mathcal{F}$ with minimal risk, only using the observations D_n, i.e. without knowing P

Table of Contents

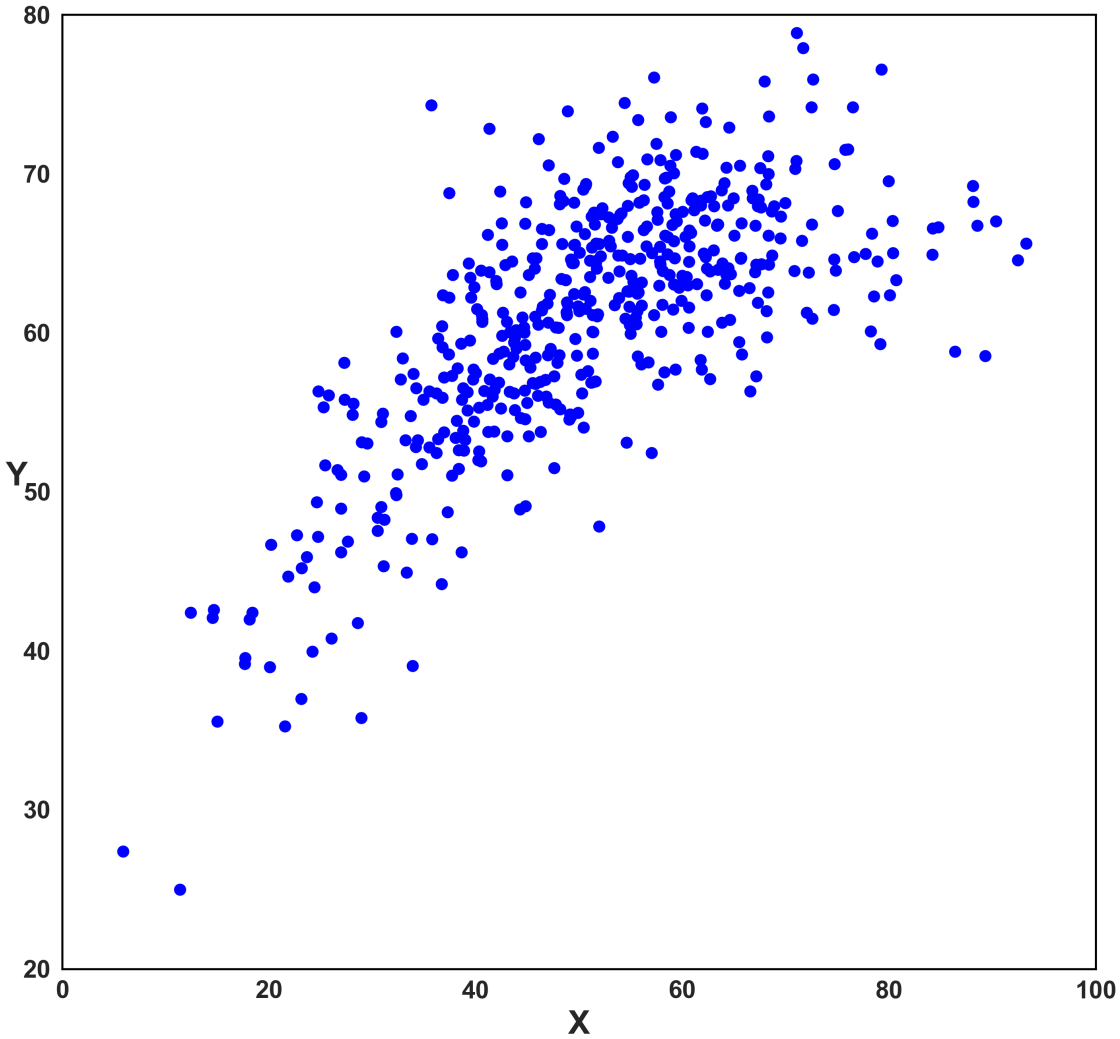
How well is it possible to predict?



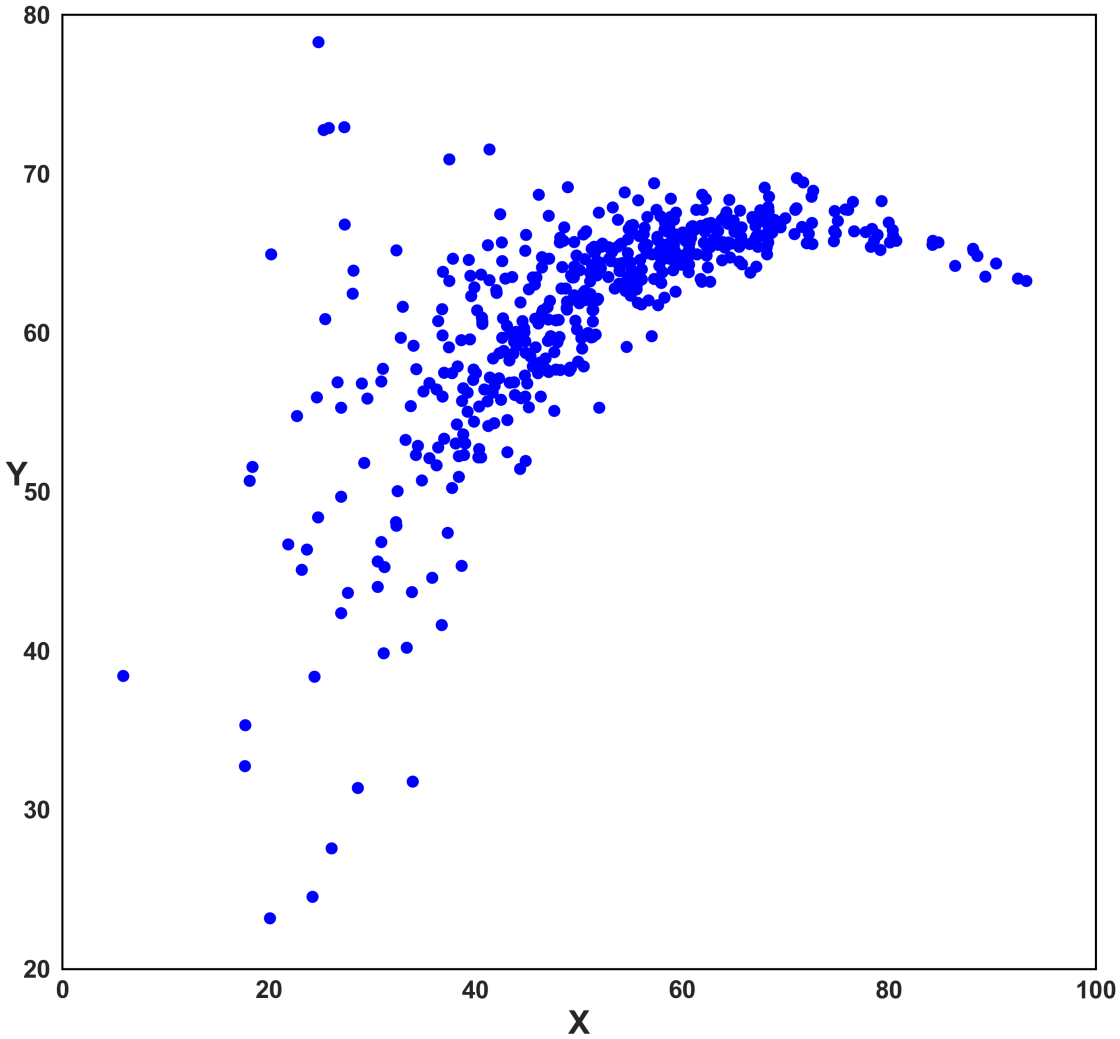
How well is it possible to predict?



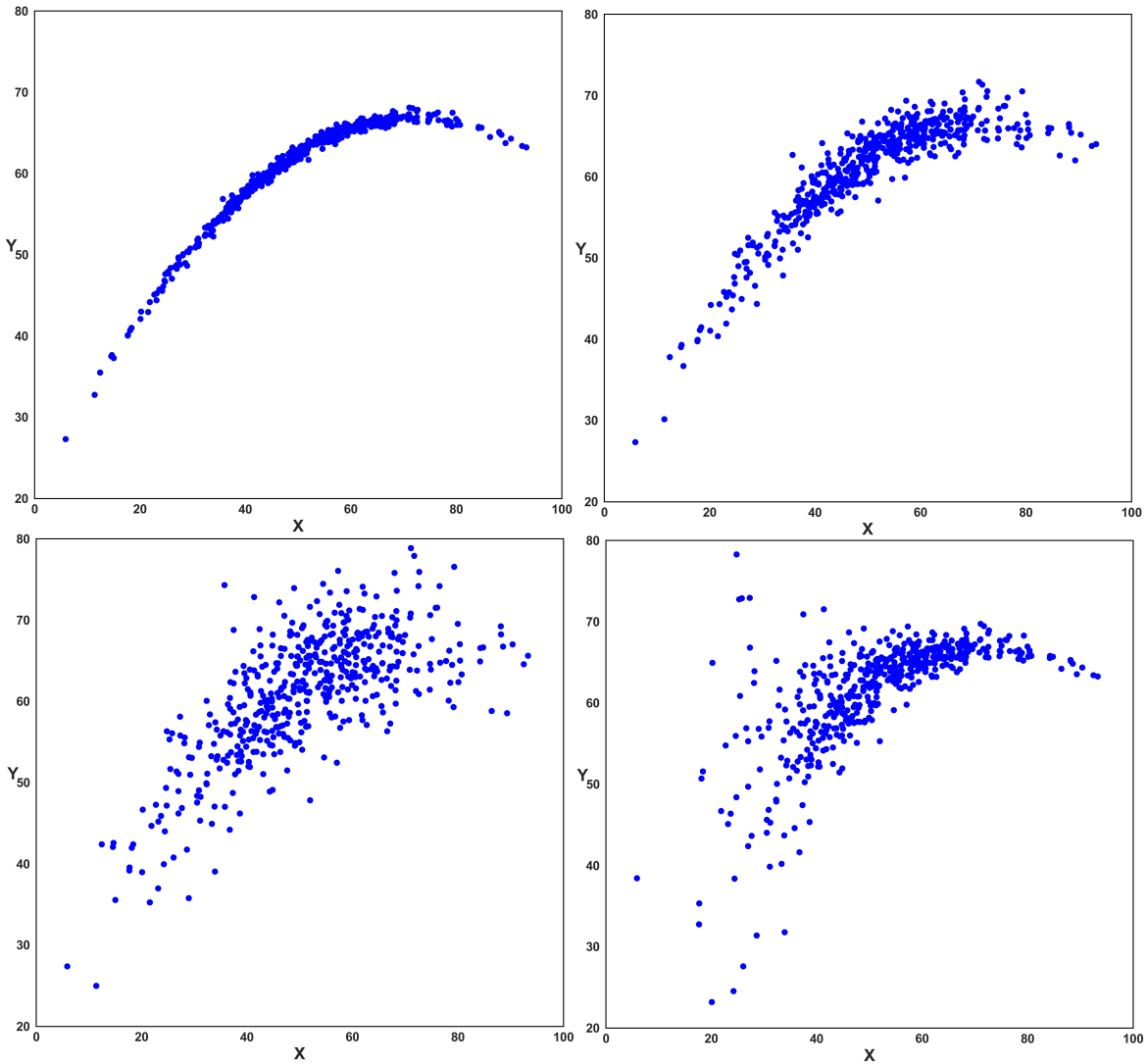
How well is it possible to predict?



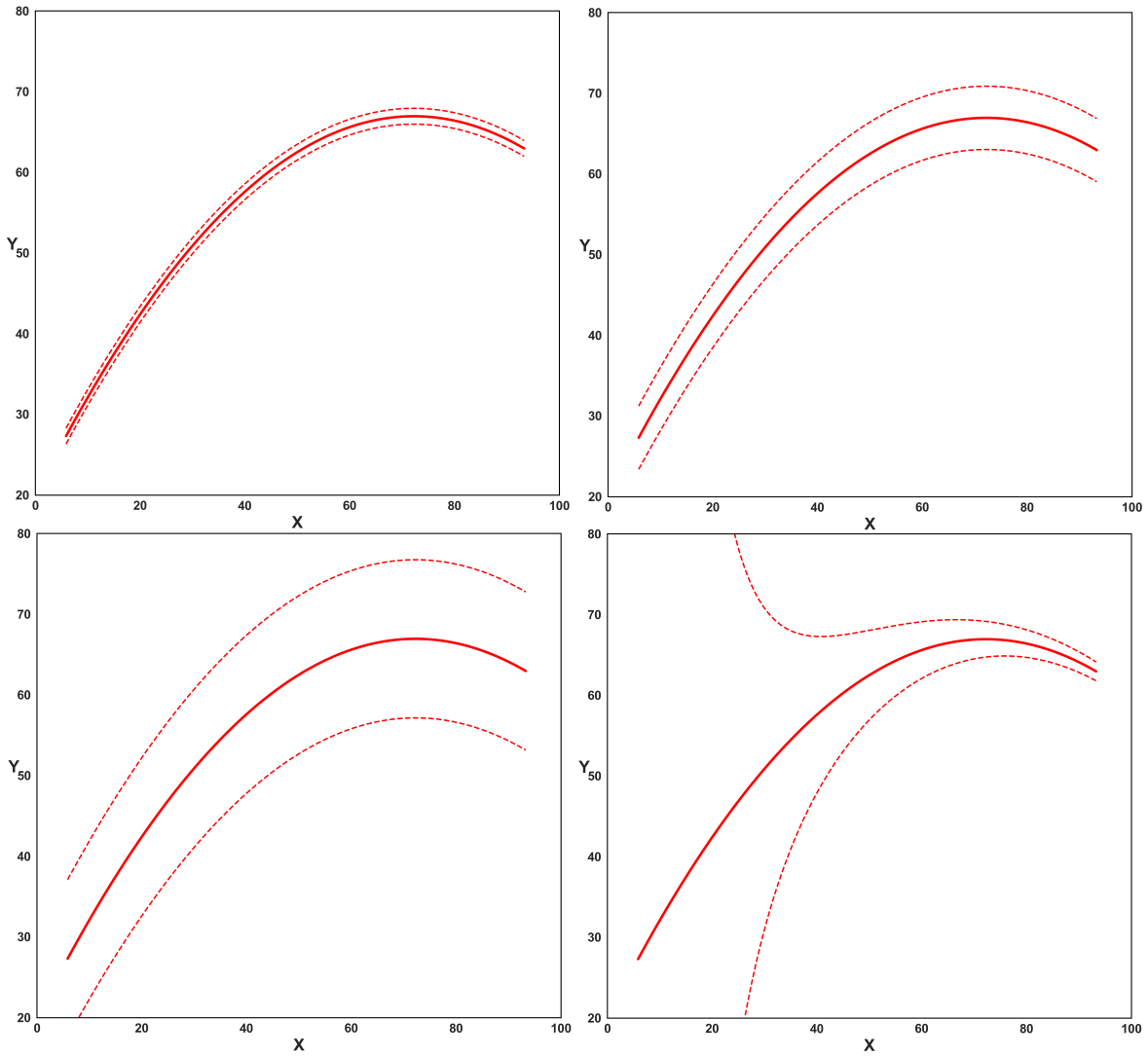
How well is it possible to predict?



How well is it possible to predict?



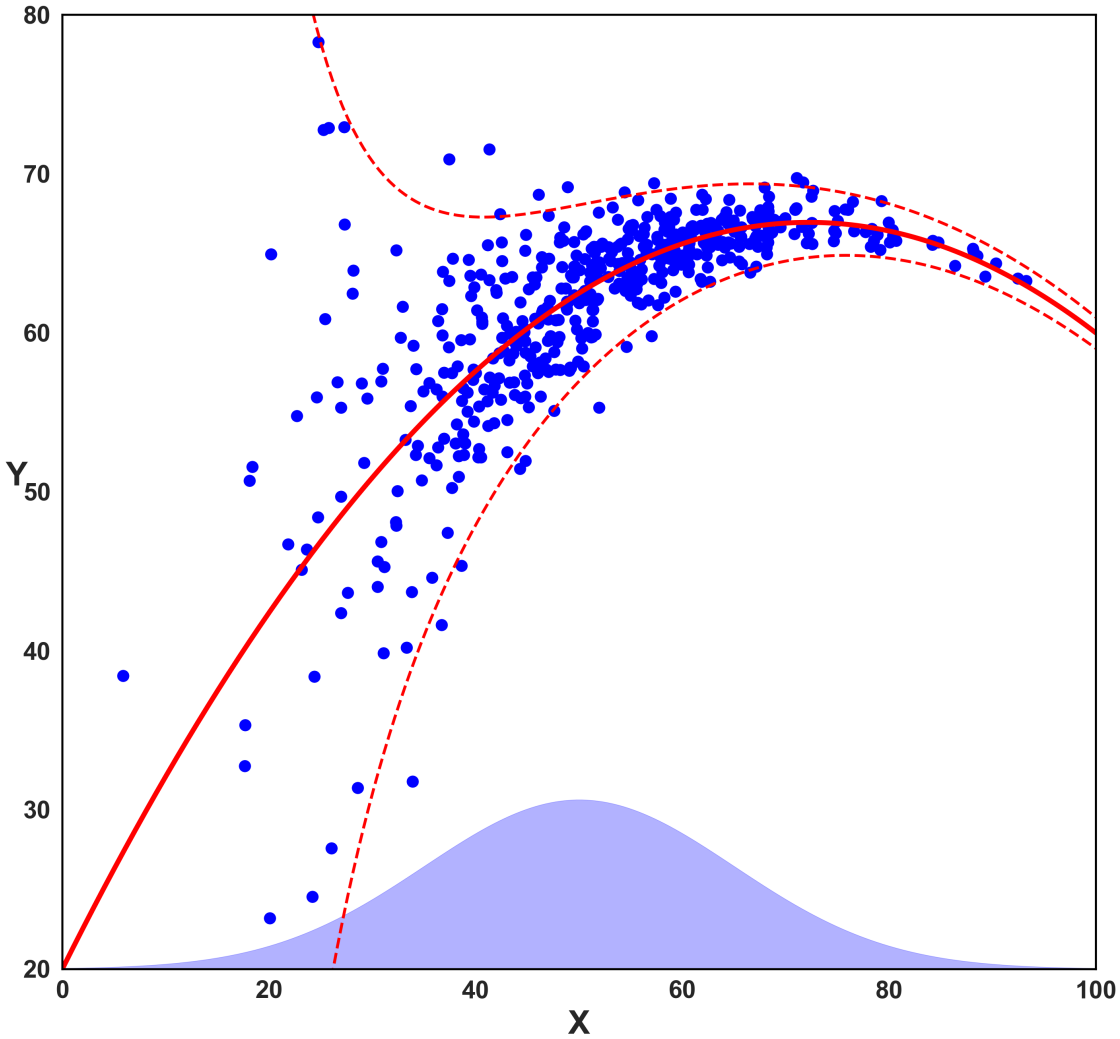
How well is it possible to predict?



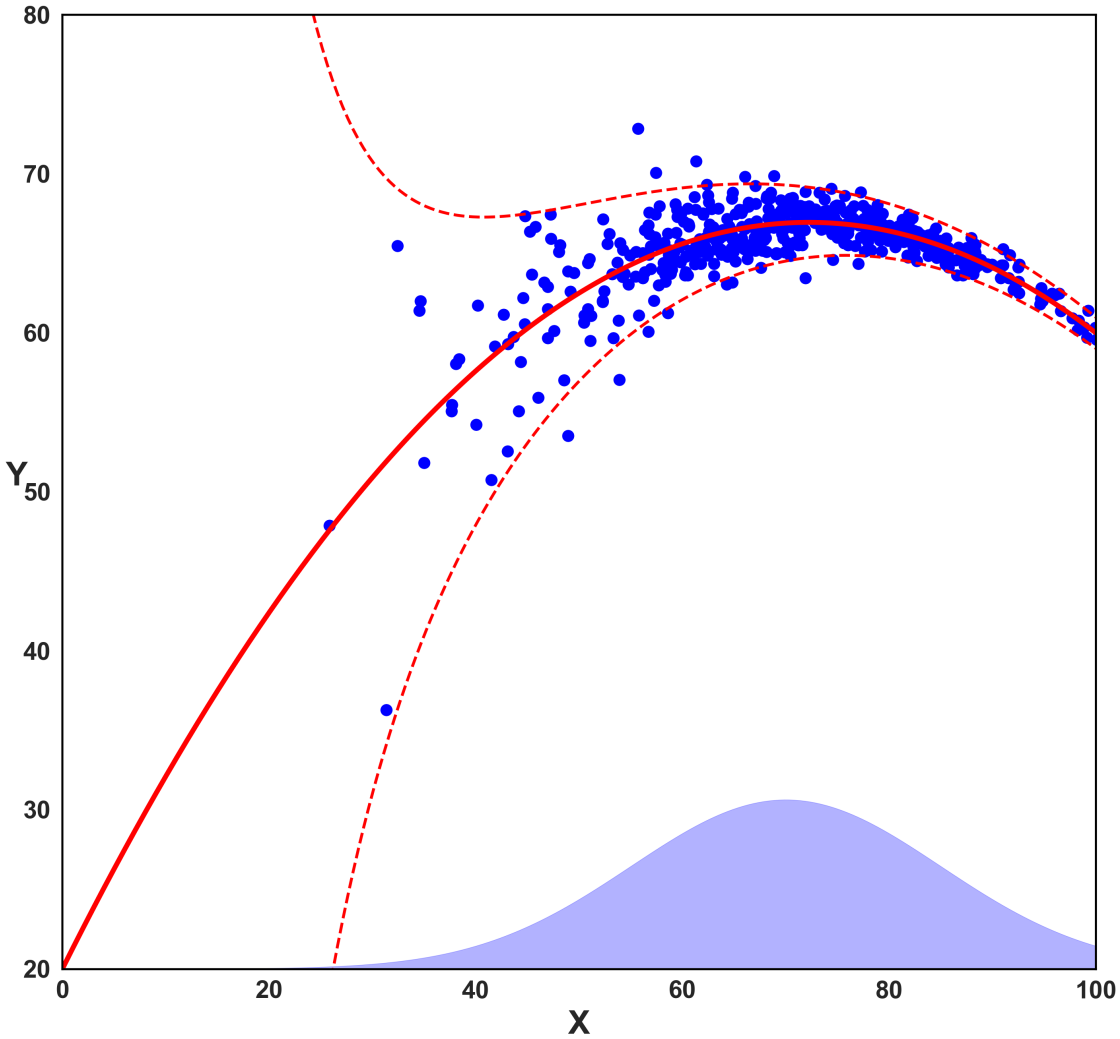
How well is it possible to predict?

- How easy or difficult it is to predict Y given X depends on the strength and complexity of their relationship
- Not all observations X are equally informative about the output Y!
- There may exists a strong relationship between certain values of X and the output. But what if observing X in this range is rare?

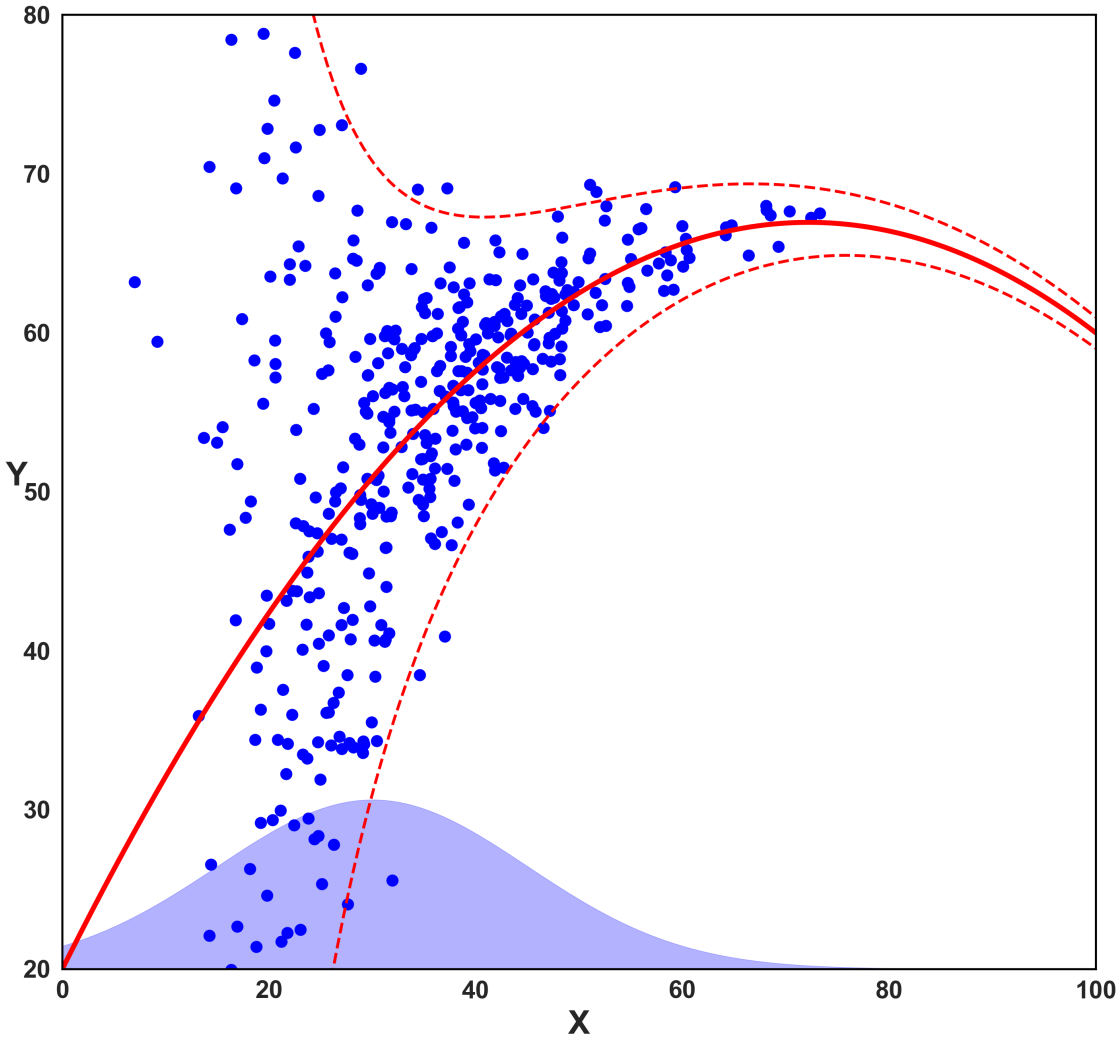
How well is it possible to predict?



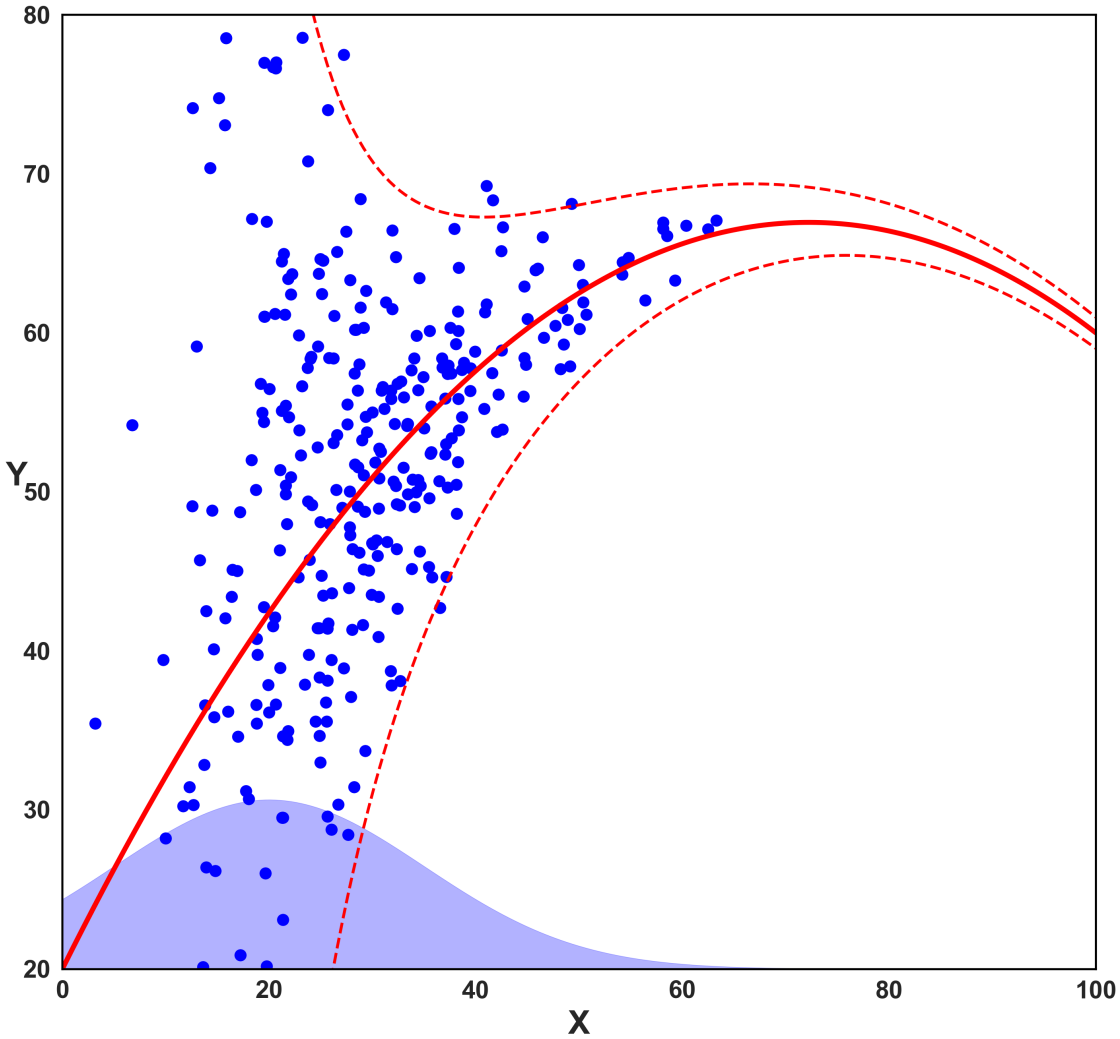
How well is it possible to predict?



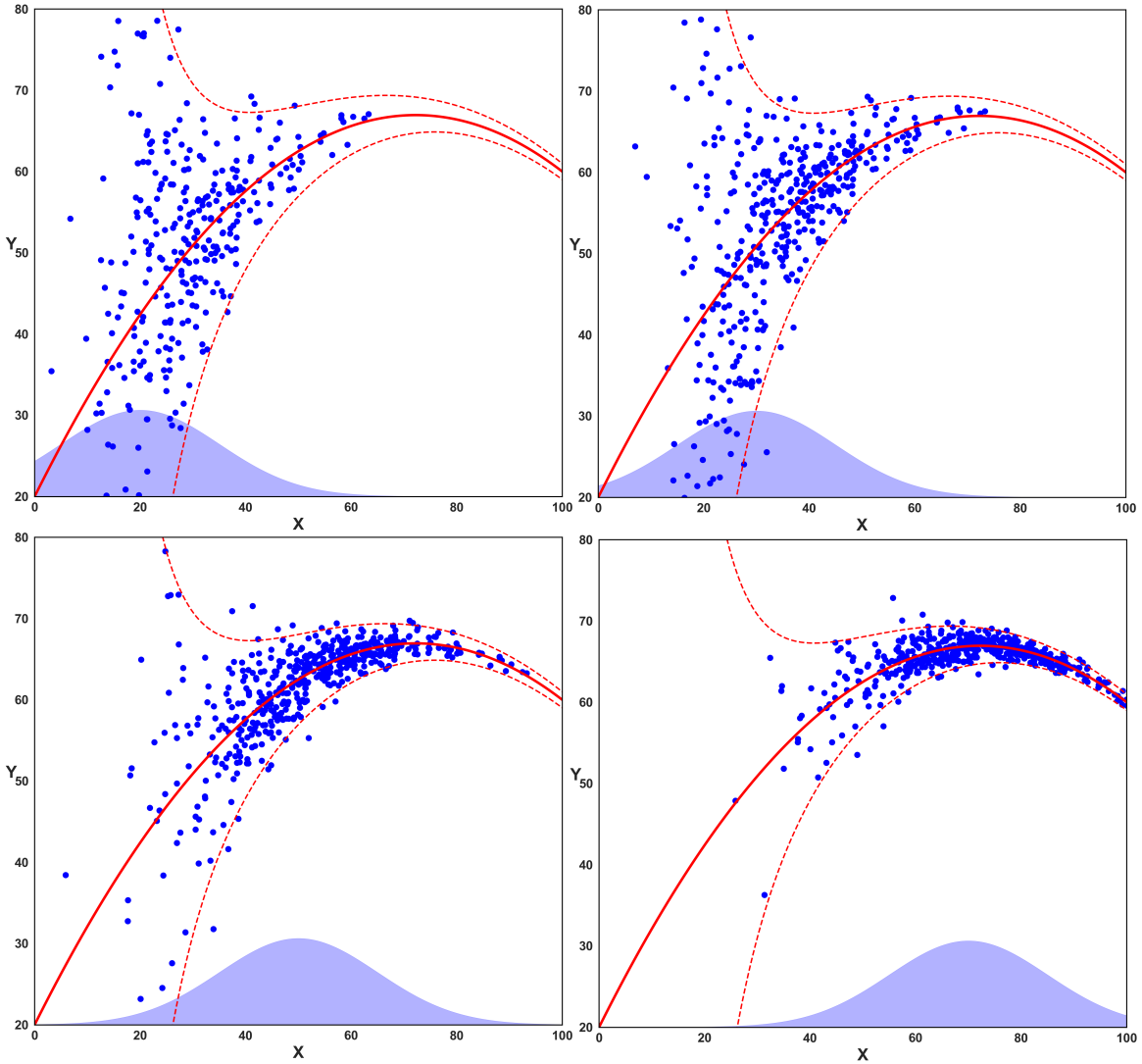
How well is it possible to predict?



How well is it possible to predict?



How well is it possible to predict?



How well is it possible to predict?

Difficulty of a learning problem

Qualitatively, for a given prediction problem, we can hope to achieve a good prediction performance if:

- the features are strongly related with the outputs in most cases
- the cases in which the observed features are uninformative occur rarely

In probability terms, we can hope to achieve a good prediction performance if for most of the likely values of X, the conditional distribution $Y|X$ is close to a deterministic function of X.

Bayes risk and target functions

IF we knew the true distribution P, we could directly try to find a predictor $f \in \mathcal{F}$ such that the risk $\mathcal{R}_P^c(f)$ is minimal: this defines Bayes risk and target functions

Bayes risk and target functions

Bayes risk is defined by

$$\mathcal{R}_P^* = \inf_{f \in \mathcal{F}} \mathcal{R}_P^c(f),$$

and any predictor $f^* \in \mathcal{F}$ such that

$$\mathcal{R}_P^c(f^*) = \mathcal{R}_P^*,$$

is called a Bayes predictor or target function. In a classification framework, we will also call such a predictor a Bayes classifier.

Bayes risk and ideal prediction

Remark

- Target functions are "ideal" predictors in the sense that their risks, i.e. *Bayes risk, is the smallest achievable by any predictor*
- Target functions are predictors *that know* the true relationship between features and labels and do not need to infer it from examples
- The Bayes risk is always well defined, but target functions do not always exist, and even if one exists, it might not be unique!

Bayes risk

Remark

The Bayes risk is

- Non-negative (infimum of a non-negative function)
- Exactly 0 only if a *perfect predictor* exists
  (technically, this can only be deduced under the assumption that a Bayes predictor exists)

In most cases, perfect predictions are unachievable and Bayes risk is then positive

Excess risk of a predictor

- When evaluating the performance of a predictor, we will be interested in its performance relative to the ideal case
- The difference between the risk of a predictor and Bayes risk is called the excess risk

Excess risk

Let $\mathcal{R}_P^*$ be the Bayes risk associated to the distribution P and cost function c. For any predictor $f \in \mathcal{F}$, we call the quantity

$$\ell(f^*, f) := \mathcal{R}_P(f) - \mathcal{R}_P^* \geq 0,$$

the excess risk of the predictor f.

Table of Contents

Setting

- Consider Y continuous and univariate, i.e. $Y \in \mathcal{Y} = \mathbb{R}$, and let c be the quadratic cost function $c(y, y') = (y - y')^2$
- The risk $\mathcal{R}_P(f) := \mathbb{E}\left[(f(X) - Y)^2\right]$ is called the *quadratic risk*

Assume $\mathbb{E}[Y^2] < +\infty$.
We can define the regression function $\eta : \mathcal{X} \to \mathbb{R}$

$$\eta(X) := \mathbb{E}[Y|X], \quad a.s.$$

(Notice that $\eta \in \mathcal{F}$ is a predictor, and assumes $P \sim (X, Y)$ known)
Then, letting $\varepsilon := Y - \eta(X)$, we can write

$$Y = \eta(X) + \varepsilon, \quad \text{with} \quad \mathbb{E}[\varepsilon|X] = 0, \quad \text{a.s.}$$

Ideal predictors in regression: how well can we possibly do?

Proposition

With $\mathcal{Y} = \mathbb{R}$, the quadratic cost, and assuming $\mathbb{E}[Y^2] < +\infty$:

1. The regression function η is a target function
2. The Bayes risk is equal to

$$\mathcal{R}_P^* = \mathbb{E}\big[(Y - \eta(X))^2\big] = \mathbb{E}\big[\mathbb{V}(Y|X)\big] = \mathbb{E}[\varepsilon^2]$$

3. The excess risk for any predictor f writes

$$\ell(f^*, f) = \mathbb{E}\big[(f(X) - \eta(X))^2\big]$$

4. A predictor $f : \mathcal{X} \to \mathbb{R}$ is a target function if and only if

$$f(X) = \eta(X), \quad \text{a.s.}$$

Interpretation

Interpretation

In the finite variance framework, the conditional expectation of Y given X, $\mathbb{E}[Y|X]$, is the best predictor possible in the sense that it minimizes the quadratic risk.

This is the reason why many parametric and non-parametric estimators attempt to estimate it from observations.

Or, you could see this result as a reason to use the quadratic cost.

Proof. (1/3)

To prove the proposition, we will adopt the following strategy:

- Show that $\mathbb{E}[\varepsilon^2]$ lower bounds the quad. risk of any predictor f
- Deduce that Bayes risk must be greater than or equal to $\mathbb{E}[\varepsilon^2]$
- Show that η, which is a predictor, achieves this lower bound
- Compute the excess risk and find predictors such that it is zero

Proof. (2/3)

For any predictor $f \in \mathcal{F}$, the quadratic risk is

$$\begin{aligned}\mathcal{R}_P(f) &= \mathbb{E}\big[(f(X)-Y)^2\big] \\ &= \mathbb{E}\big[(f(X)-\eta(X)-\varepsilon)^2\big] \\ &= \mathbb{E}\big[(f(X)-\eta(X))^2\big] - 2\underbrace{\mathbb{E}\big[\varepsilon(f(X)-\eta(X))\big]}_{=\mathbb{E}\left[(f(X)-\eta(X))\mathbb{E}[\varepsilon|X]\right]=0} + \mathbb{E}[\varepsilon^2].\end{aligned}$$

For any $f \in \mathcal{F}$, we thus have $\mathcal{R}_P(f) \geq \mathbb{E}[\varepsilon^2]$.
Hence,

$$\mathcal{R}_P^* \overset{def.}{=} \inf_{f\in\mathcal{F}} \mathcal{R}_P(f) \geq \mathbb{E}[\varepsilon^2].$$

Moreover, we notice that $\mathcal{R}_P(\eta) = \mathbb{E}[\varepsilon^2]$, and therefore $\mathcal{R}_P^* = \mathcal{R}_P(\eta) = \mathbb{E}[\varepsilon^2]$, showing points 1 and 2.

Proof. (3/3)

Now, by definition of the excess risk of a predictor f:

$$\begin{aligned}\ell(f^*, f) &\overset{def.}{=} \mathcal{R}_P(f) - \mathcal{R}_P^* \\ &= \mathbb{E}\big[(f(X) - \eta(X))^2\big] + \mathbb{E}[\varepsilon^2] - \mathbb{E}[\varepsilon^2] \\ &= \mathbb{E}\big[(f(X) - \eta(X))^2\big].\end{aligned}$$

This shows point 3.
Finally, a predictor f is a target function iff $\ell(f^*, f) = 0$.

$$\ell(f^*, f) = 0 \Leftrightarrow \mathbb{E}\big[(f(X) - \eta(X))^2\big] = 0 \Leftrightarrow (f(X) - \eta(X))^2 = 0 \text{ a.s.},$$

$(f(X) - \eta(X))^2 \geq 0$

which is equivalent to $f(X) = \eta(X)$ a.s., showing point 4. □

Table of Contents

Setting

- Consider $Y \in \mathcal{Y} = \{0, 1\}$
  Ex: absence/presence of a pedestrian on an image
- Let c be the 0-1 cost function $c(y, y') = \mathbb{1}_{y \neq y'}$
  $\mathcal{R}_P(f) := \mathbb{E}\big[\mathbb{1}_{f(X) \neq Y}\big] = \mathbb{P}\big(f(X) \neq Y\big)$ is called the *0-1 risk*

Define again the function $\eta : \mathcal{X} \to \mathbb{R}$ such that

$$\eta(X) := \mathbb{E}[Y|X] = \mathbb{P}\big(Y = 1 \big| X\big), \quad a.s.$$

Notice here that η is not exactly a predictor because
$\eta : \mathcal{X} \to \mathbb{R} \not\subset \mathcal{Y}$, but is the conditional likelihood of "label=1"

Ideal predictors in classif.: how well can we possibly do?

Proposition

With $\mathcal{Y} = \{0, 1\}$ and the 0-1 cost:

1. The predictor $x \mapsto f^*(x) := \mathbb{1}_{\eta(x)>1/2}$ is a target function
2. The Bayes risk is equal to
$$\mathcal{R}_P^* = \mathbb{E}\Big[\min\big\{\eta(X), 1-\eta(X)\big\}\Big]$$
3. The excess risk for any predictor f writes
$$\ell(f^*, f) = \mathbb{E}\Big[\big|2\eta(X) - 1\big|\mathbb{1}_{f^*(X)\neq f(X)}\Big]$$
4. A predictor $f : \mathcal{X} \to \{0, 1\}$ is a target function iff
$$f(X) = \mathbb{1}_{\eta(X)>1/2}, \quad \text{a.s.},$$
except, possibly, on the event $\{\eta(X) = 1/2\}$.

Interpretation

Interpretation

The predictor f^* basically predicts the most likely of the two outcomes.

Proof. (1/6)

To prove the proposition, we will adopt a similar strategy to that in the regression case:

- Show that $\mathbb{E}\big[\min\big\{\eta(X), 1-\eta(X)\big\}\big]$ lower bounds the 0-1 risk of any predictor f
- Deduce that Bayes risk must be greater than or equal to this lower bound
- Show that f^* achieves this lower bound
- Compute the excess risk and find predictors such that it is zero

Proof. (2/6)

For any predictor $f \in \mathcal{F}$, consider the 0-1 risk conditionally on X:

$$r_X(f) := \mathbb{P}\big(f(X) \neq Y \big| X\big).$$

$$r_X(f) = \mathbb{P}\big(\{f(X) = 0 \text{ and } Y = 1\} \text{ or } \{f(X) = 1 \text{ and } Y = 0\}\big|X\big)$$

σ-additivity

$$= \mathbb{P}\big(\{f(X) = 0 \text{ and } Y = 1\}\big|X\big) + \mathbb{P}\big(\{f(X) = 1 \text{ and } Y = 0\}\big|X\big)$$

Also, $f(X)$ is constant given X.
Hence, $f(X)$ and Y are independent given X, and we have

$$r_X(f) = \mathbb{P}\big(f(X) = 0\big|X\big)\mathbb{P}\big(Y = 1\big|X\big) + \mathbb{P}\big(f(X) = 1\big|X\big)\mathbb{P}\big(Y = 0\big|X\big)$$

Proof. (3/6)

Thus:

$$\begin{aligned} r_X(f) &= \mathbb{1}_{f(X)=0}\mathbb{P}\big(Y=1\big|X\big) + \mathbb{1}_{f(X)=1}\mathbb{P}\big(Y=0\big|X\big) \\ &= \mathbb{1}_{f(X)=0}\eta(X) + \mathbb{1}_{f(X)=1}(1-\eta(X)) \\ &\geq \min\big\{\eta(X), 1-\eta(X)\big\} \end{aligned}$$

By the law of iterated expectations we have:

$$\mathbb{E}[r_X(f)] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{f(X)\neq Y}|X\right]\right] = \mathbb{E}\left[\mathbb{1}_{f(X)\neq Y}\right] = \mathcal{R}_P(f)\,,$$

and thus, taking expectations on both sides of the inequality above, we have that for any predictor f:
$\mathcal{R}_P(f) \geq \mathbb{E}\Big[\min\big\{\eta(X), 1-\eta(X)\big\}\Big]$, and hence

$$\mathcal{R}_P^* \overset{def.}{=} \inf_{f\in\mathcal{F}} \mathcal{R}_P(f) \geq \mathbb{E}\Big[\min\big\{\eta(X), 1-\eta(X)\big\}\Big].$$

Proof. (4/6)

We obtained a lower bound for Bayes risk.
Let's show that this bound is reached for $f = f^*$
(recall $x \mapsto f^*(x) := \mathbb{1}_{\eta(x)>1/2}$). Notice that

$$\begin{aligned} r_X(f^*) &= \mathbb{1}_{f^*(X)=0}\eta(X) + \mathbb{1}_{f^*(X)=1}(1-\eta(X)) \\ &= \min\big\{\eta(X), 1-\eta(X)\big\}. \end{aligned} \qquad \text{(Show it)}$$

Hence, taking expectations on both sides as before:

$$\mathbb{P}\big(f^*(X) \neq Y\big) = \mathcal{R}_P(f^*) = \mathbb{E}\Big[\min\big\{\eta(X), 1-\eta(X)\big\}\Big].$$

Therefore: $\mathcal{R}_P(f^*) = \mathcal{R}_P^* = \mathbb{E}\Big[\min\big\{\eta(X), 1-\eta(X)\big\}\Big]$,
which shows points 1 and 2.

Proof. (5/6)

Let's now show points 3 and 4.
Recall we found that for any predictor f, the 0-1 risk given X writes:

$$\mathbb{P}\big(f(X) \neq Y \big| X\big) \overset{def}{=} r_X(f) = \mathbb{1}_{f(X)=0}\eta(X) + \mathbb{1}_{f(X)=1}(1-\eta(X)).$$

Thus (Show the following)

$$\begin{aligned} r_X(f) = \quad & \mathbb{1}_{f(X)=f^*(X)} \min\big\{\eta(X), 1-\eta(X)\big\} \\ & + \mathbb{1}_{f(X)\neq f^*(X)} \max\big\{\eta(X), 1-\eta(X)\big\}, \end{aligned}$$

[Hint: $f(X) \neq f^*(X) \Leftrightarrow f(X) = 1 - f^*(X)$]

and

$$\begin{aligned} & r_X(f) - r_X(f^*) \\ & \quad = \mathbb{1}_{f(X)\neq f^*(X)} \Big[\max\big\{\eta(X), 1-\eta(X)\big\} - \min\big\{\eta(X), 1-\eta(X)\big\} \Big], \end{aligned}$$

Proof. (6/6)

Taking expectations on both sides and using that: $\forall a, b \in \mathbb{R}$,

$$\max\{a, b\} - \min\{a, b\} = |a - b|,$$

yields the excess risk

$$\mathcal{R}_P(f) - \mathcal{R}_P^* \overset{def}{=} \ell(f^*, f) = \mathbb{E}\Big[\big|2\eta(X) - 1\big| \mathbb{1}_{f(X) \neq f^*(X)}\Big].$$

This shows point 3. Finally we obtain point 4:
$f \in \mathcal{F}$ is a target function iff $\ell(f^*, f) = 0$, i.e.

$$\mathbb{E}\Big[\big|2\eta(X) - 1\big| \mathbb{1}_{f(X) \neq f^*(X)}\Big] = 0 \iff \big|2\eta(X) - 1\big| \mathbb{1}_{f(X) \neq f^*(X)} = 0 \text{ a.s.}$$
$$\iff \{\eta(X) = 1/2 \text{ or } f(X) = f^*(X)\} \text{ a.s. } \square$$

Summary

- We formalised the learning problem:
  Notions of samples, feature/label relationship, predictors, risk of prediction.
- Defined prediction performance benchmarks:
  Notions of ideal predictors (target functions/Bayes predictors), best performance possible (Bayes risk).
- These ideal predictors *know* the true distribution/relationship between features and labels.
- We obtained the Bayes risk and corresponding Bayes predictor for two fundamental frameworks:
  - Regression with quadratic cost
  - Binary classification with 0-1 cost

Assignment

- You can now make questions 1, 2 and 3 of Assignment Part I
- I encourage you to already start working on it (deadline of Part I February 29th at 23:59)