Introduction

Statistical learning theory: Inferring quantities, functions, class assignments, ... using data and assumptions about its distribution.

Making inferences from training data, together with assumptions about its distribution, for the purpose of making predictions about test (new) data.

1. Learning from data

Point estimation: Estimate a population parameter θ of a random observable X from sample data $x_1, \ldots, x_p$ and assumptions about their distribution.

Examples:

(1) Find an estimate, $\widehat{\theta}$ of the average height of an adult German from a sample $x_1, \ldots, x_p$ of the heights of p randomly selected individuals.

(1′) Find an estimate $\widehat{\theta}_f$ (resp., $\widehat{\theta}_m$) of the average height of a German female (resp., male) from a sample $x_1, \ldots, x_p$ of the heights (not genders!) of p randomly selected individuals under the assumption that the heights of females and males are each normally distributed. This is an *incomplete data problem*.

Function estimation: Find a plausible relationship $Y \approx \widehat{f}(X)$ between random observables X and Y from data $(x_1, y_1), \ldots, (x_p, y_p)$ and assumptions about their distribution. When Y is continuous (resp., discrete) this is typically called *regression* (resp., classification).

Examples:

(1) Let X and Y be the height and weight of a randomly selected adult German female, respectively. Find the "best" approximation to Y as a function of X assuming
- X and Y are jointly normal.
- $Y|X$ is normal with constant variance.
- nothing.

(2) Armed with a large training set $(x_1, y_1), \ldots, (x_p, y_p)$ where x_j is an image and $y_j = 1$ if x_j contains a cat and $y_j = 0$ otherwise, find a function

$$\widehat{f} : \{\text{images}\} \longrightarrow \{0, 1\}$$

such that if $x \neq x_j$ is an image, then $\widehat{f}(x)$ is likely to be 1 if x contains a cat and is likely to be 0, otherwise.

(3) Call an image x an *occlusion* of an image y if x made from y by overlaying it with a small white patch. Given a large training set $(x_1, y_1), \ldots, (x_p, y_p)$ where y_j is an image and x_j is an occlusion of y_j, find a function

$$\widehat{f} : \{\text{images}\} \longrightarrow \{\text{images}\}$$

such that if $y \neq y_j$ is an image of and x is an occlusion of y, then $\widehat{f}(x) \approx y$.

Questions:

(1) How can we find good point/function estimates?

(2) How can we compare different point/function estimates? Is there a “best” estimate?