

## Unit 1

# Introduction and overview

# Introduction

# Statistical and machine learning

- Machine learning is an umbrella term for solving problems by machines “discovering” their “own” algorithms.
- Statistical learning theory is a framework for machine learning drawing from the fields of statistics and functional analysis.
- Statistical learning theory deals with the problem of finding a predictive function based on data.
- The goals of learning are understanding and prediction.

# Terminology

- One distinguishes between:
  - Supervised learning:
    - Outcome measure available.
    - Regression, classification.
  - Unsupervised learning:
    - No outcome measure.
    - Clustering.
- The variables have different roles assigned:
  - Input: features, regressors, covariate, independent variable.
  - Output: outcome, dependent variable.
- Training set of data.
- The predictive model is a *learner*.

# Overview

# Overview of supervised learning

- One assumes that

$$Y = f(X) + \epsilon.$$

- One aims at determining  $\hat{f}$  which is the estimate of  $f$  to predict  $Y$  based on a training data set:

$$\hat{Y} = \hat{f}(X).$$

# Overview of supervised learning / 2

- The accuracy of  $\hat{Y}$  as a prediction of  $Y$  depends on two quantities:
  - ➊ Reducible error.
  - ➋ Irreducible error.

$$E[(Y - \hat{Y})^2] = E[(f(X) + \epsilon - \hat{f}(X))^2] = (f(X) - \hat{f}(X))^2 + \text{Var}(\epsilon),$$

where  $(f(X) - \hat{f}(X))^2$  is the reducible error and  $\text{Var}(\epsilon)$  is the irreducible error.

- The reducible error can be further split into
  - ➊ Bias
  - ➋ Variance

depending on the mean and variance of  $\hat{f}$  if it is repeatedly estimated using a large number of training data sets.

# Overview of supervised learning / 3

- One differentiates between:
  - Regression: quantitative output.
  - Classification: qualitative output.
- Two simple approaches to prediction:
  - Least squares:
    - Huge assumptions.
    - Stable, but possibly inaccurate predictions.
  - Nearest neighbors:
    - Mild structural assumptions.
    - Accurate, but unstable predictions.



# Linear models and least squares

- Given a vector of inputs  $X^T = (X_1, X_2, \dots, X_p)$  we predict the output  $Y$  via the model

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^p X_j \hat{\beta}_j.$$

- If the intercept / the constant variable 1 is included in  $X$ , then one can write

$$\hat{Y} = X^T \hat{\beta}.$$

- In the  $(p + 1)$ -dimensional input-output space  $(X, \hat{Y})$  represents a hyperplane. If the constant is included in  $X$ , the hyperplane goes through the origin and is a subspace. Otherwise it is an affine set.

# Linear models and least squares / 2

- Viewed as a function over the  $p$ -dimensional input space

$$f(X) = X^T \beta$$

is linear and the gradient

$$f'(X) = \beta$$

is a vector in input space that points in the steepest uphill direction.

- The least squares fit is obtained by minimizing

$$\text{RSS}(\beta) = \sum_{i=1}^N (y_i - \mathbf{x}_i^T \beta)^2 = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

with respect to  $\beta$  given  $N$  observations.

# Linear models and least squares / 3

- $\text{RSS}(\beta)$  is a quadratic function of the parameters:
  - There always exists a minimum.
  - The minimum might not be unique.
- If  $\mathbf{X}^T \mathbf{X}$  is nonsingular, the unique solution is given by:

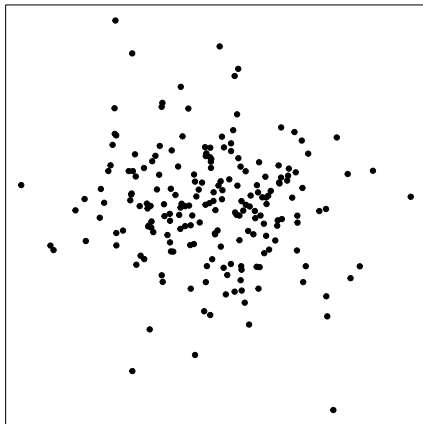
$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

- The fitted and predicted values are given by:

$$\hat{y}_i = \mathbf{x}_i^T \hat{\beta}$$

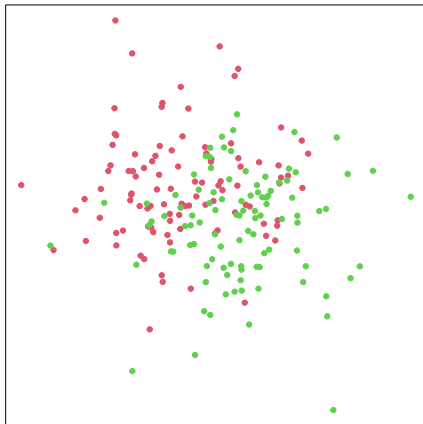
$$\hat{y}(x_0) = \mathbf{x}_0^T \hat{\beta}.$$

# Example



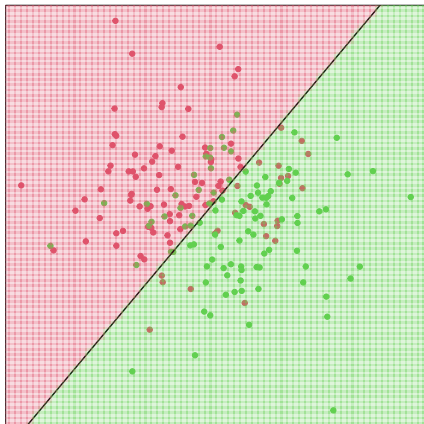
## Example / 2

Data points with true class labels



# Example / 3

Linear regression of 0/1 response



# Nearest neighbor method

- The observations in the training set  $\mathcal{T}$  closest in input space to  $x$  are used to form the prediction  $\hat{Y}$  at position  $x$ .
- The  $k$  nearest neighbor fit for  $\hat{Y}$  is defined as:

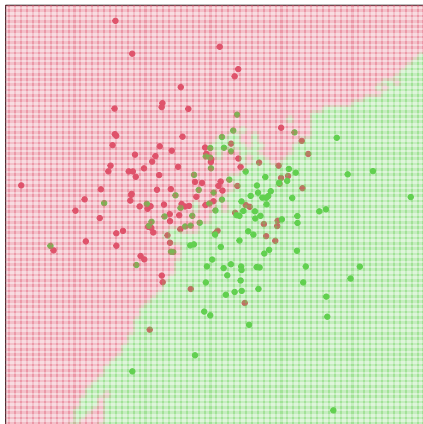
$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i,$$

where  $N_k(x)$  is the neighborhood of  $x$  defined by the  $k$  closest points  $x_i$  in the training sample.

- Closeness implies a metric. For now, Euclidean distance.
- For  $k$  nearest neighbor fits the error on the training data should be approximately an increasing function of  $k$ , and will always be 0 for  $k = 1$ .
- In  $k$  nearest neighbors there is one parameter,  $k$ , to choose:
  - Effective number of parameters:  $\frac{N}{k}$ .
  - Minimizing the error in the training set would always choose  $k = 1$ .

# Example

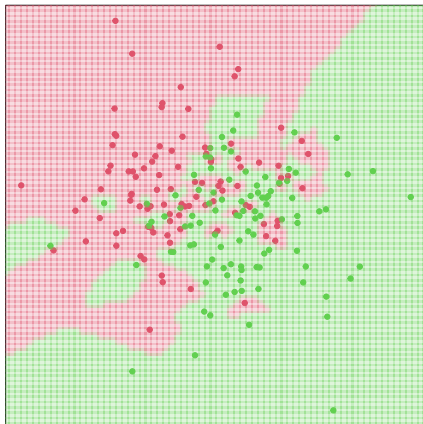
20 nearest neighbor classifier





# Example / 2

1 nearest neighbor classifier



# From least squares to nearest neighbors

- Least squares leads to a linear decision boundary:
  - Smooth and apparently stable to fit.
  - Appears to heavily rely on the assumption that a linear decision boundary is appropriate.

⇒ Low variance, potentially high bias.
- $k$  nearest neighbors method
  - Relies on no stringent assumptions about the underlying data.
  - Any particular subregion of the decision boundary depends only on a handful of input points.

⇒ High variance, low bias.

# Statistical decision theory

- Given:
  - Assume we have a quantitative output.
  - Let  $X \in \mathbb{R}^p$  denote a real valued random input vector.
  - Let  $Y \in \mathbb{R}$  denote a real valued random output variable.
  - Denote the joint distribution by  $\Pr(X, Y)$ .
- Target:
  - Determine a function  $f(X)$  for predicting  $Y$  which minimizes a loss function

$$L(Y, f(X)),$$

which penalizes prediction errors.

# Statistical decision theory / 2

- Possible loss functions are for example the *squared error loss*

$$L(Y, f(X)) = (Y - f(X))^2.$$

- This leads to the expected squared prediction error as criterion for choosing  $f$ :

$$\text{EPE}(f) = \mathbb{E}(Y - f(X))^2 = \int [y - f(x)]^2 \Pr(dx, dy).$$

- By conditioning on  $X$  we obtain

$$\text{EPE}(f) = \mathbb{E}_X \mathbb{E}_{Y|X}([Y - f(X)]^2 | X),$$

using  $\Pr(X, Y) = \Pr(Y|X)\Pr(X)$ .

# Statistical decision theory / 3

- This implies that it suffices to minimize EPE pointwise:

$$f(x) = \arg \min_c E_{Y|X}([Y - c]^2 | X = x),$$

with the solution

$$f(x) = E_{Y|X}(Y | X = x).$$

This is the conditional expectation, also known as *regression* function.

# Statistical decision theory: $k$ nearest neighbors

- Nearest neighbor methods attempt to estimate this conditional mean using

$$\hat{f}(x) = \text{Ave}(y_i | x_i \in N_k(x)),$$

where  $\text{Ave}(\cdot)$  denotes average and two approximations are exploited:

- ❶ Expectation is approximated by averaging over the sample data.
- ❷ Conditioning at a point is relaxed to conditioning on some region “close” to the target point.

# Statistical decision theory: $k$ nearest neighbors / 2

- Thus, under mild regularity conditions on the joint probability distribution  $\Pr(X, Y)$ , one can show that as  $N, k \rightarrow \infty$  such that  $k/N \rightarrow 0$ ,

$$\hat{f}(x) \rightarrow E(Y|X = x).$$

- Problems:
  - If only a small sample size is available, exploiting structure in the data could lead to more stable estimators.
  - The *rate* of convergence depends on the dimension of the feature space. The rate decreases if the dimension increases.

# Statistical decision theory: linear regression

- Assumes that the regression function  $f(x)$  is approximately linear in its arguments:

$$f(x) \approx x^T \beta.$$

- Thus, we specify a model for the regression function and minimize the EPE. This gives

$$\beta = [E(XX^T)]^{-1}E[XY].$$

- Linear regression does not condition on  $X$ , but pools over all values of  $X$  using the knowledge of the functional relationship.
- Estimating  $\beta$  consists of replacing the expectation by taking the average over the training data.



# Statistical decision theory: comparison

- Model assumptions:
  - Least squares assumes  $f(x)$  is well approximated by a globally linear function.
  - $k$  nearest neighbors assumes  $f(x)$  is well approximated by a locally constant function.

# Statistical decision theory: loss functions

- So far we considered the squared error loss, also referred to as  $L_2$  loss.
- Alternatives are for example the  $L_1$  loss:

$$L(Y, f(X)) = |Y - f(X)|,$$

which if this expected loss is minimized gives the conditional median as estimate:

$$\hat{f}(x) = \text{median}(Y|X = x).$$

- $L_1$  criteria have discontinuities in their derivatives, which have hindered their widespread use.

# Loss functions: categorical outcome

- The categorical outcome  $G$  takes values in  $\mathcal{G}$ , the set of possible classes.
- The estimate  $\hat{G}$  also takes values in  $\mathcal{G}$ .
- The loss can be represented by a  $K \times K$  matrix  $\mathbf{L}$ , where  $K = \text{card}(\mathcal{G})$ .
- The matrix  $\mathbf{L}$  has
  - zero values on the diagonal,
  - nonnegative values elsewhere,

where  $L(l, k)$  denotes the price to pay for classifying an observation belonging to the  $k$ th class  $\mathcal{G}_k$  into the  $l$ th class  $\mathcal{G}_l$ .

- The EPE can be written as

$$\text{EPE}(\hat{G}) = \mathbb{E}(L(G, \hat{G}(X))) = \mathbb{E}_X \left[ \sum_{k=1}^K L(\mathcal{G}_k, \hat{G}(X)) \Pr(\mathcal{G}_k | X) \right].$$

# Loss functions: categorical outcome / 2

- The pointwise minimization is again sufficient:

$$\hat{G}(x) = \arg \min_{g \in \mathcal{G}} \sum_{k=1}^K L(\mathcal{G}_k, g) \Pr(\mathcal{G}_k | X = x).$$

- Examples for loss functions:
  - Zero-one loss function:

$$L(l, k) = \begin{cases} 0 & l = k, \\ 1 & l \neq k. \end{cases}$$

The EPE is minimized for the zero-one loss by

$$\hat{G}(x) = \arg \min_{g \in \mathcal{G}} [1 - \Pr(g | X = x)] = \max_{g \in \mathcal{G}} \Pr(g | X = x).$$

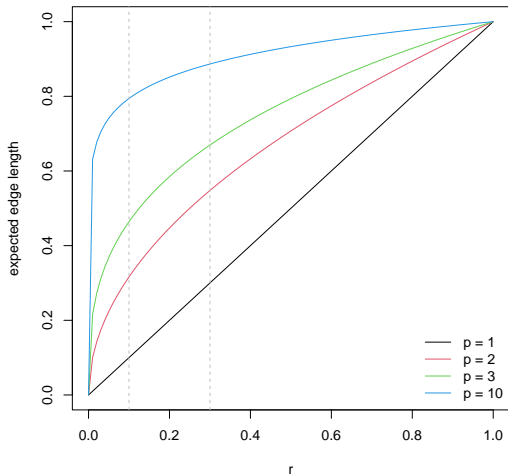
This solution is called *Bayes classifier*. The error rate of the Bayes classifier is called *Bayes rate*.

# Local methods in high dimensions

- High dimensional problems suffer from the *curse of dimensionality* (Bellman, 1961).
  - Observations tend to have no “close” neighbors.
  - Observations are closer to the boundary than to any other data point.
- Example: Assume uniformly distributed data in the  $p$ -dimensional unit cube. If the fraction  $r$  of observations to be contained in a hypercubical neighborhood is fixed, the expected edge length of this cube is given by

$$e_p(r) = r^{1/p}.$$

# Local methods in high dimensions / 2



# Model selection and the bias-variance tradeoff

- Many statistical learning models contain a *smoothing* or *complexity* parameter.
- More complex models will in general have a better performance on the training data, but this will not translate to a better performance on new test data.
- The expected squared prediction error at point  $x_0$  for a fixed procedure to estimate  $f$  by  $\hat{f}$  based on the training set  $\mathcal{T}$  is given by

$$\begin{aligned}\text{EPE}_{\hat{f}_{\mathcal{T}}}(x_0) &= \text{E}[(Y - \hat{f}_{\mathcal{T}}(x_0))^2 | X = x_0] \\ &= \text{Var}(Y | X = x_0) + \text{E}_{\mathcal{T}}[(\text{E}(Y | X = x_0) - \hat{f}_{\mathcal{T}}(x_0))^2 | X = x_0] \\ &= \text{Var}(Y | X = x_0) + [\text{Bias}_{\mathcal{T}}^2(\hat{f}_{\mathcal{T}}(x_0)) + \text{Var}_{\mathcal{T}}(\hat{f}_{\mathcal{T}}(x_0))]\end{aligned}$$

# Model selection and the bias-variance tradeoff / 2

- This decomposition indicates:
  - The first term is the *irreducible* error.
  - The second and third terms are the *mean squared error* of  $\hat{f}_{\mathcal{T}}(x_0)$  in estimating  $f(x_0)$  decomposed into *bias* and *variance*.
- In general one has:
  - If the model complexity is increased, the bias is reduced.
  - If the model complexity is increased, the variance is increased.



# Model selection and the bias-variance tradeoff / 3

