Lecture 2

Agenda

- **Mathematical Formulation - A First Go**
  - Binary classification - Goal and Probabilistic setup
  - The Principle of Empirical Risk Minimization (ERM)
  - Concentration Bounds - McDarmid's Inequality
  - Complexity (Combinatorial) - VC Dimension
- **Some Popular Classifications Methods - Heuristics**
  - Parametric Approach: Linear Logistic Regression
  - The (single layer) Perceptron Algorithm
  - K-Nearest Neighbours
  - Decision Trees - The CART Algorithm
- **Assessing the Accuracy of the Results**
  - Cross Validation
  - Bootstrap - The Plug-in Principle

Probabilistic setup for binary classification

- Random pair $= (X, Y) \sim P$ unknown
- $X =$ observation vector in $\mathcal{X}$ (ex: $\mathbb{R}^d$ with $d >> 1$)
- $Y =$ binary label in $\mathcal{Y} = \{-1, +1\}$
- **Our goal:** guess the *output* Y from the *input* observation X
- **Classifier**: $C : x \in \mathcal{X} \mapsto C(x) \in \{-1, 1\}$ in a **class** $\mathcal{G}$
- Risk functional (unknown!) $=$ **Expected prediction error**

$$L(C) = \mathbb{E}[Y \neq C(X)]$$

to minimize over $C \in \mathcal{G}$.

Theoretical Risk Minimization

- Let $\eta(x) = \mathbb{P}(Y = +1 | X = x)$ **regression function**
- Let $p = \mathbb{P}(Y = +1)$
- Compute $C^* = \arg\min_{C \in \mathcal{G}} L(C)$
- Calculations yields the **Naive Bayes Classifier**

$$C^*(x) = 2 \cdot \mathbb{I}\{\eta(x) > 1/2\} - 1, \quad x \in \mathcal{X}$$

$\Rightarrow$ affects the likeliest label given the observation $X = x$

- Minimum theoretical risk: $L^* = L(C^*) = 1/2 - \mathbb{E}[|\eta(X) - 1/2|]$
- How close $\eta(X)$ is to $1/2$ governs the difficulty of the problem!

Theoretical Risk Minimization

- Theoretical **excess of risk**:

$$L(C) - L^* = \mathbb{E}[|\eta(X) - 1/2|\mathbb{I}\{X \in G^* \Delta G_C\}]$$

where G^*, G_C denote the subsets of the input space $\mathcal{X}$

$$\begin{aligned} G^* &= \{\eta(X) > 1/2\} \\ G_C &= \{C(X) = +1\} \end{aligned}$$

and $A\Delta B = (A \cap \bar{B}) \cup (\bar{A} \cap B)$ the *symmetric difference*.

- Insights: when a little of X's mass is concentrated around the **margin** $\{\eta(x) = 1/2\}$, the problem gets simpler.

Empirical Risk Minimization (ERM)

- Data $= D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$
- Classifier candidate: $C : \mathcal{X} \to \{-1, 1\}$ in a class $\mathcal{G}$
- Empirical risk functional $=$ Training (misclassification) error

$$L_n(C) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i \neq C(X_i)\}$$

  to minimize over $C \in \mathcal{G}$.

- Solution **"empirical risk minimizer"**: $\hat{C}_n = \arg\min_{C \in \mathcal{G}} L_n(C)$
- OK for the training data, now for **future data** (X, Y)?

Investigating the properties of the ER Minimizer

- Don't forget that $\hat{C}_n$ is **random** (depending on the data D_n)
- Let $(X, Y) \sim P$ be a **new random pair**, independent from D_n
  Will $\hat{C}_n$ performs well as a classifier for this novel pair?

$$\Rightarrow \text{ compute } L(\hat{C}_n) = \mathbb{P}(Y \neq \hat{C}_n(X) \mid D_n)$$

- $L(\hat{C}_n)$ is a **random variable**! It depends on the data D_n.
- **Deviation** between the r.v. $L(\hat{C}_n)$ and the min. error L^* (cst)

$$\Rightarrow \text{ Study the excess of risk } 0 \leq \mathcal{E}(C) = L(\hat{C}_n) - L^*$$

- Learning Theory: compute explicit **confidence bounds**, $\forall \epsilon > 0$

$$\mathbb{P}_{D_n}(L(\hat{C}_n) - L^* \geq \epsilon) \leq ?$$

Learning Bounds

- Consider $C_0 = \arg\min_{C \in \mathcal{G}} L(C)$ (theoret. minimizer over $\mathcal{G}$)
- Check the **"bias-variance"** decomposition

$$L(\hat{C}_n) - L^* \quad \leq \quad 2 \sup_{C \in \mathcal{G}} |L(C) - \hat{L}_n(C)| + L(C_0) - L^*$$

- The second term depends on the model $\mathcal{G}$ solely (bias)
- The 1st term (estimation) involves **concentration** of

$$Z = \{L(C) - \hat{L}_n(C)\}_{C \in \mathcal{G}}$$

$\Rightarrow$ theory of **empirical processes**

Empirical processes - Basics

- Let $X_1, \ldots, X_n$ be i.i.d. r.v.'s drawn as P
- Let $P_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ the empirical df
- Let $\mathcal{F}$ be a class of functions $f : \mathbb{R} \to \mathbb{R}$
- **Empirical process** $\{P_n f\}_{f \in \mathcal{F}}$: $P_n f = n^{-1} \sum_{i=1}^{n} f(X_i)$, $f \in \mathcal{F}$
- Investigate which conditions on $\mathcal{F}$ allow to **control**

$$||Z|| = \sup_{f \in \mathcal{F}} |P_n f - P f|$$

- Ex.: recall **Donsker's theorem**, $\mathcal{F} = \{\mathbb{I}\{. \leq x\},\ x \in \mathbb{R}\}$

$$\sqrt{n} \sup_{x \in \mathbb{R}} |n^{-1} \sum_{i \leq n} \mathbb{I}\{X_i \leq x\} - P(]-\infty, x])| \Rightarrow \sup_{t \in [0,1]} |B(t)|$$

Basics inequalities

- Finite class: $Card(\mathcal{F}) = N$.

  "Union's bound" combined with **Chernoff's method**

$$\mathbb{P}(\sup_{f \in \mathcal{F}} |P_n f - Pf| \geq \epsilon) \leq 2N \cdot e^{-2n\epsilon^2}$$

  if $\forall f \in \mathcal{F} : \ 0 \leq f \leq 1$

- Cumulative distribution functions: **Dvoretsky-Kiefer-Wolfowitz**

$$\mathbb{P}(\sqrt{n} \sup_{x \in \mathbb{R}} |\frac{1}{n} \sum_{i \leq n} \mathbb{I}\{X_i \leq x\} - P(]-\infty, x])| \geq \epsilon) \leq 2e^{-2\epsilon^2}$$

- McDarmid (1989)

Measuring Complexity - Combinatorial Approach

- Vapnik - Chervonenkis: **VC dimension** of a class $\mathcal{A}$ of subsets $A \subset \mathbb{R}^d$
- Let $x_1^n = (x_1, \ldots, x_n)$ be n points in $\mathbb{R}^d$. Define
  - **Trace:**
$$Tr(\mathcal{A}, x_1^n) = \{A \cap x_1^n;\ A \in \mathcal{A}\}$$
  - **Shattering coefficient:**
$$S_{\mathcal{A}}(n) = \max_{x_1^n} Card\, Tr(\mathcal{A}, x_1^n)$$
  - Ex: half-lines of $\mathbb{R}$: $S_{\mathcal{A}}(n) = n + 1$
- Other approaches: entropy metric, Rademacher chaos, *etc.*

Parametric approach - Parametric logistic regression

- Explicit modelling of $\eta(x) = \mathbb{P}(Y = +1 \mid X = x) \in]0, 1[$
- **Logistic** transform: $f(x) = \operatorname{logit} \eta(x) = \log(\frac{\eta(x)}{1-\eta(x)})$
- Inverse transform: $\eta(x) = \frac{e^{f(x)}}{1+e^{f(x)}}$
- Assume $f \in \mathcal{F} = \{f_\theta(x);\ \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$

$$\eta_\theta(x) = \frac{e^{f_\theta(x)}}{1 + e^{f_\theta(x)}}$$

- Ex: **linear** logistic regression $f(x) = \alpha +^t \beta \cdot x, \quad \theta = (\alpha, \beta)$
- Maximize the **log-likelihood**

$$l_n(\theta;\ x_1, \ldots, x_n) = \sum_{i=1}^{n} \{y_i \log(\eta_\theta(x_i)) + (2y_i - 1) \log(1 - \eta_\theta(x_i))\}$$

The (single-layer) perceptron algorithm

- The output Y is connected to the input X by

$$y = sign({}^t w \cdot X - \beta)$$

- The input space is separated into two regions by a **hyperplane**
- **Rosenblatt's algorithm (1962)** for minimizing

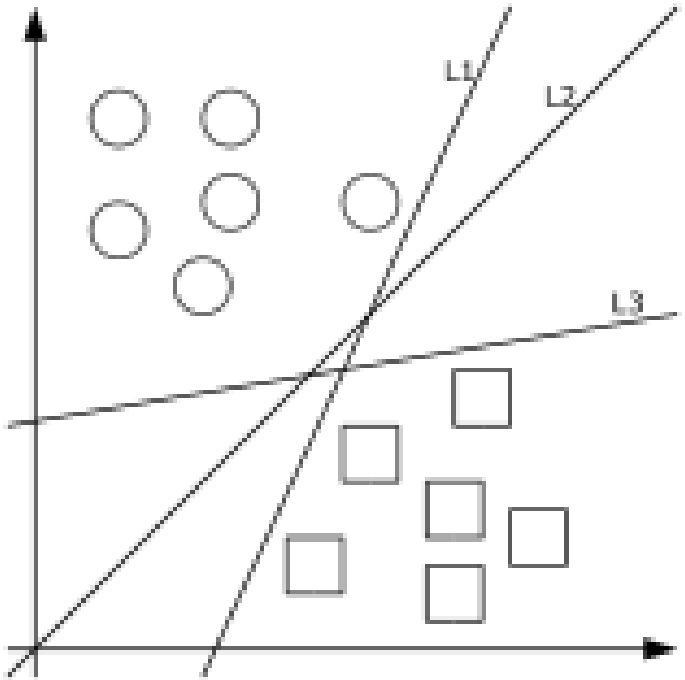
$$-\sum_i y_i({}^t w \cdot x_i + \beta)$$

1. Choose at random (x_i, y_i) for "feeding" the perceptron
2. Gradient descent with rate ρ

$$\begin{pmatrix} w \\ \beta \end{pmatrix} \leftarrow \begin{pmatrix} w \\ \beta \end{pmatrix} + \rho \begin{pmatrix} y_i x_i \\ y_i \end{pmatrix}$$

3. Converges only when the data are **separable in a linear fashion**

The (single-layer) perceptron algorithm



A simplistic nonparametric method: K-nearest neighbours

- Let $K \geq 1$. On $\mathbb{R}^D$, consider a **metric** d (ex: euclidean distance)
- For any input value x, let $\sigma = \sigma_x$ be the permutation of $\{1, \ldots, n\}$ such that

$$d(x, x_{\sigma(1)}) \leq \ldots \leq d(x, x_{\sigma(n)})$$

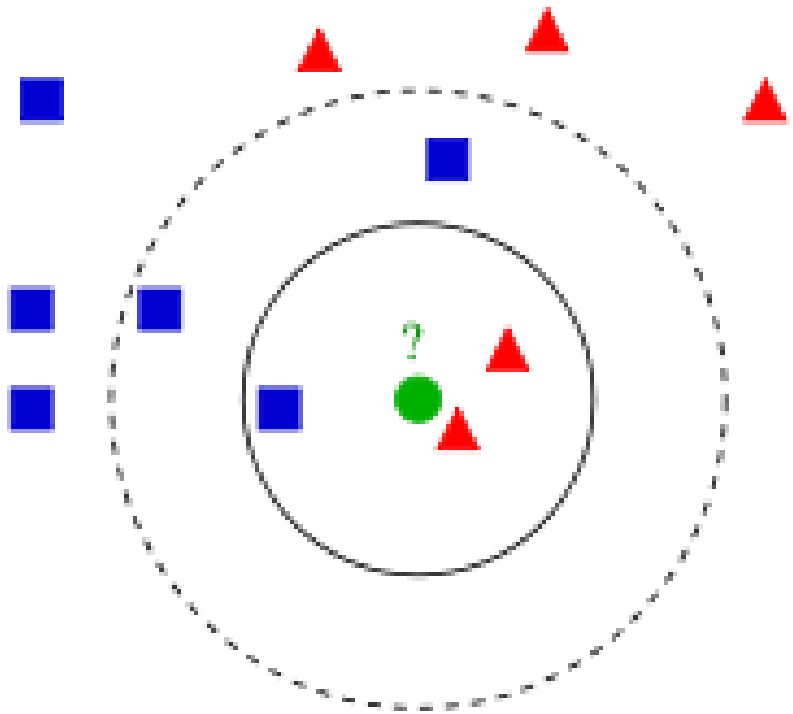
- Consider the K-**nearest neighbours**

$$\{x_{\sigma(1)}, \ldots, x_{\sigma(K)}\}$$

- **Majority vote:** $N_y = \text{Card}\{k \in \{1, ..., K\};\ y_{\sigma(k)} = y\}$, $y \in \{-1, 1\}$

$$C(x) = \arg \max_{y \in \{-1, +1\}} N_y,$$

A simplistic nonparametric method: K-nearest neighbours



K-nearest neighbours

Consistency (Stone '77)

If $k = k_n \to \infty$ such that $k_n = o(n)$, then the K-NN rule is consistent

$$L(C_{K-NN}) - L^* \to 0, \quad as \quad n \to \infty$$

But...

- The rate can be arbitrarily slow
- Instability: choice of K? metric D?

Decision Trees: the CART Algorithm

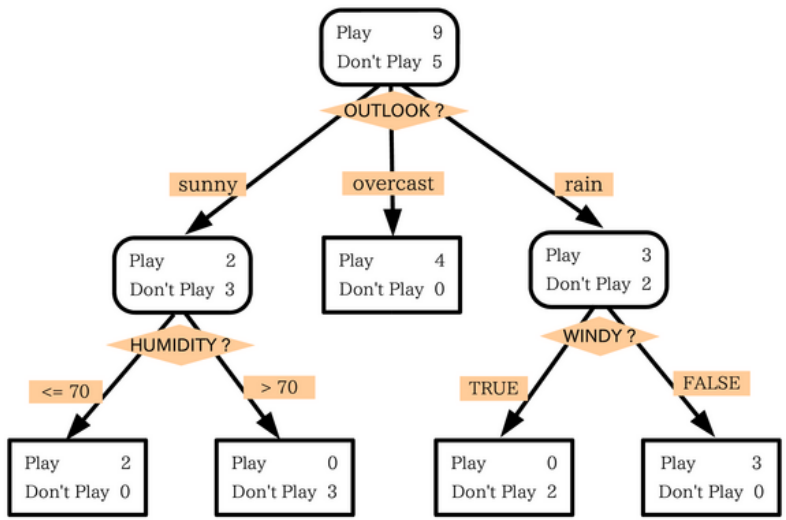
- Breiman, Friedman, Olshen & Stone (1986)
- Recursive Dyadic Partitioning: $X = (X^{(1)}, \ldots, X^{(d)}) \in \mathbb{R}^d$
- "Growing the Tree": iterate
  1. For $j = 1$ to d, find s (best split value) so as to minimize the impurity of the regions
  $$\{X_j > s\} \ \textit{and} \ \{X_j \leq s\}$$
  2. Find the best split variable X_j
- Measuring **impurity**:
  - misclassification error
  - Gini index

Decision Trees: the CART Algorithm

Play golf dataset

Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

Dependent variable: PLAY



Validation

- When data are not expensive: **cross-validation**

  Training - Test - Validation

- Bootstrap (the plug-in principle): estimate the distribution of

$$\mathbb{E}^*[\mathbb{I}\{\hat{C}(X) \neq Y\}]$$

where $\mathbb{E}^*[.]$ is the expectation w.r.t. the empirical df of the $(X_i, Y_i)'s$