# Lecture
# –
# Basic algorithms for pattern recognition

# 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**

$$l_n(\theta) = \sum_{i=1}^{n} \left\{ \frac{1+y_i}{2} \log(\eta_\theta(x_i)) + \frac{1-y_i}{2} \log(1 - \eta_\theta(x_i)) \right\}$$

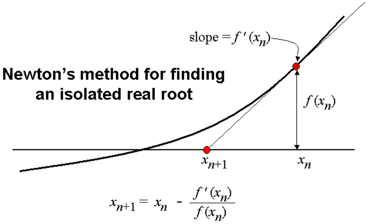
# Parametric approach - Parametric logistic regression

- Even in the additive model, the score equation

$$\nabla_\theta l_n(\theta) = 0$$

cannot be solved explicitly!

- Implement Newton-Raphson method (gradient descent)

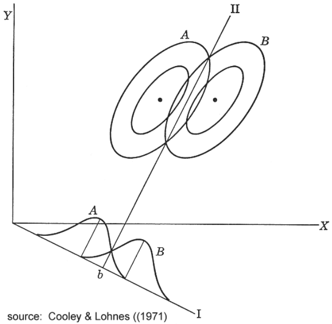



- Alternative to logit: probit model $\Phi^{-1}(\eta(X)) = \alpha +^t \beta X$

$$\text{with } \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\frac{t^2}{2}) dt.$$

# Parametric approach - Linear Discriminant Analysis

- Assume that the conditional distributions of $X$ given $Y = +1$ and given $Y = -1$ are **Gaussian** with same covariance matrix $\Gamma$ but different means $\mu_+$ and $\mu_-$. Let $p = \mathbb{P}\{Y = +1\}$.
- Estimate the moments of first and second orders, next the likelihood ratio and assign the likeliest labels




source: Cooley & Lohnes ((1971)

Figure 9.1.

# Parametric approach - Linear Discriminant Analysis

At point $X$, predict $Y = +1$ if

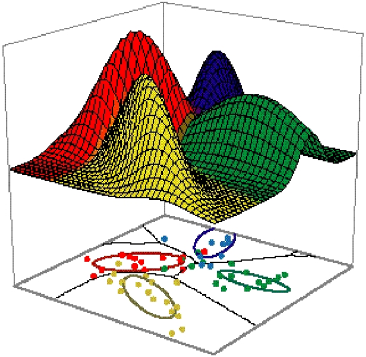
$$\log\left(\frac{\mathbb{P}\{Y = +1 \mid X\}}{\mathbb{P}\{Y = -1 \mid X\}}\right) > 0 \Leftrightarrow$$

$$\log(\frac{p}{1-p}) - \frac{1}{2}(\mu_+ - \mu_-)^t\Gamma^{-1}(\mu_+ - \mu_-) + x^t\Gamma^{-1}(\mu_+ - \mu_-) > 0$$

- Linear separator ($\neq$ linear logistic regression, except when $p = 1/2$)
- Replace $\mu_+$, $\mu_-$ and $\Gamma$ by empirical estimates

# Parametric approach - Linear Discriminant Analysis

- Naive Bayes: given $Y$, the input variables $X^{(1)}, \ldots, X^{(d)}$ are independent
- Nonlinear decision boundaries: quadratic discriminant analysis (QDA), Gaussian mixtures, kernels
- LDA can be easily extended to the multiclass framework



# The (single-layer) perceptron algorithm

- The output $Y$ is connected to the input $X$ by

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

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

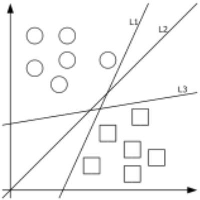
$$-\sum_i y_i({}^t w \cdot x_i + \theta)$$

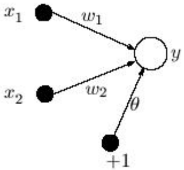
1. Choose at random $(x_i, y_i)$ for "feeding" the perceptron
2. Gradient descent with rate $\rho$

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

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

# The (single-layer) perceptron algorithm

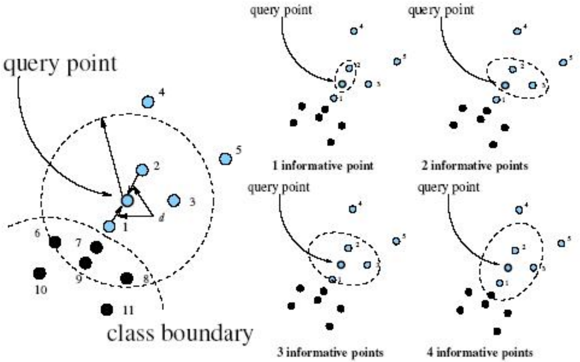


# A simplistic nonparametric method: $K$-nearest neighbours

- 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:** $N_y = \mathrm{Card}\{k \in \{1, ..., K\};\ y_{\sigma(k)} = y\}$, $y \in \{-1, 1\}$
$$C(x) = \arg \max_{y \in \{-1,+1\}} N_y,$$

# A simplistic nonparametric method: $K$-nearest neighbours



# $K$-nearest neighbours

**Consistency (Stone '77)**

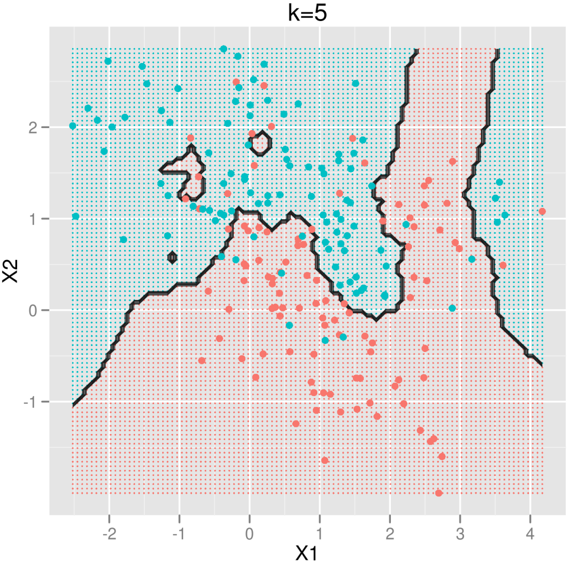
If $k = k_n \to \infty$ such that $k_n = o(n)$, then the $K$-NN rule is consistent

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

But...

- The rate can be arbitrarily slow
- Curse of dimensionality: sorting data is computationally expensive
- Instability: choice of $K$? metric $D$?
- **Metric learning** (*e.g.* Mahalanobis distance)
- Variants with *weights*

# $K$-nearest neighbours - A too flexible method?



# Histogram rules - Local averaging

- K-NN limitations: a nearest neighbor may be very far from $X$!
- Consider a **partition** of the feature space:

$$C_1 \bigcup \cdots \bigcup C_K = \mathcal{X}$$

- Apply the **majority rule**: suppose that $X$ lies in $C_k$,
  1. Count the number of training examples with positive label lying in $C_k$
  2. If $\sum_{i:\ X_i \in C_k} \mathbb{I}\{Y_i = +1\} > \sum_{i:\ X_i \in C_k} \mathbb{I}\{Y_i = -1\}$, predict $Y = +1$. Otherwise predict $Y = -1$.
- This corresponds to the "plug-in" classifier $2\mathbb{I}\{\widehat{\eta}(x)\} - 1$, where

$$\widehat{\eta}(x) = \sum_{k=1}^{K} \mathbb{I}\{x \in C_k\} \frac{\sum_{i=1}^{n} \mathbb{I}\{Y_i = +1,\ X_i \in C_k\}}{\sum_{i=1}^{n} \mathbb{I}\{X_i \in C_k\}}$$

is the **Nadaraya-Watson estimator** of the posterior probability.

# Kernel rules - Local averaging

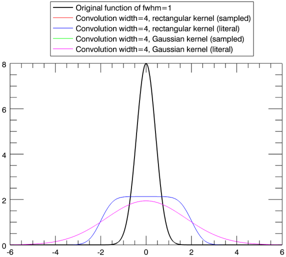
- Smooth the estimator/boundary decision!
- Replace the indicator function by a **convolution kernel**:

$$K : \mathbb{R}^d \longrightarrow \mathbb{R}_+, \quad K \geq 0, \text{ symmetric and } \int K(x)dx = 1$$

- Bandwidth $h > 0$ and **rescaling**

$$K_h(x) = \frac{1}{h} K(x/h)$$

- Examples: Gaussian kernel, Novikov, Haar, *etc.*



# Kernel rules - Local averaging

- If $\sum_{i=1}^{n} \mathbb{I}\{Y_i = +1\} K_h(x - X_i) > \sum_{i=1}^{n} \mathbb{I}\{Y_i = -1\} K_h(x - X_i)$, predict $Y = +1$. Otherwise predict $Y = -1$.
- This corresponds to the "plug-in" classifier $2\mathbb{I}\{\widetilde{\eta}(x)\} - 1$, where

$$\widetilde{\eta}(x) = \frac{\sum_{i=1}^{n} \mathbb{I}\{Y_i = +1\} K_h(x - X_i)}{\sum_{i=1}^{n} K_h(x - X_i)}$$

  is the **Nadaraya-Watson estimator** of the posterior probability.
- **Statistical argument:** if $\eta$ is a "smooth" function, $\widetilde{\eta}$ may be a better estimate than $\widehat{\eta}$ (smaller variance but... biased)