MITx: Statistics, Computation & Applications

Genomics and High-Dimensional Data Module

Lecture 2: Classification with Hig-Dimensional Data

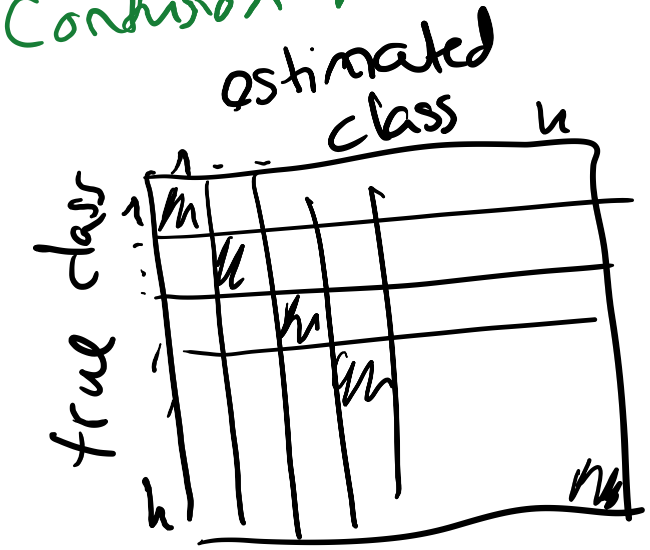
$x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^p$

$c^{(1)}, \dots, c^{(n)} \in \mathcal{C}$

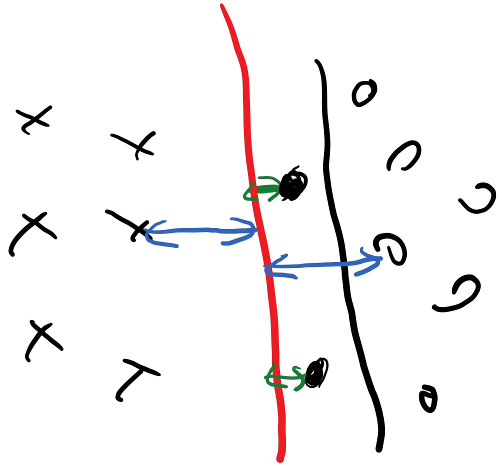
$f: \mathbb{R}^p \longrightarrow \mathcal{C}$

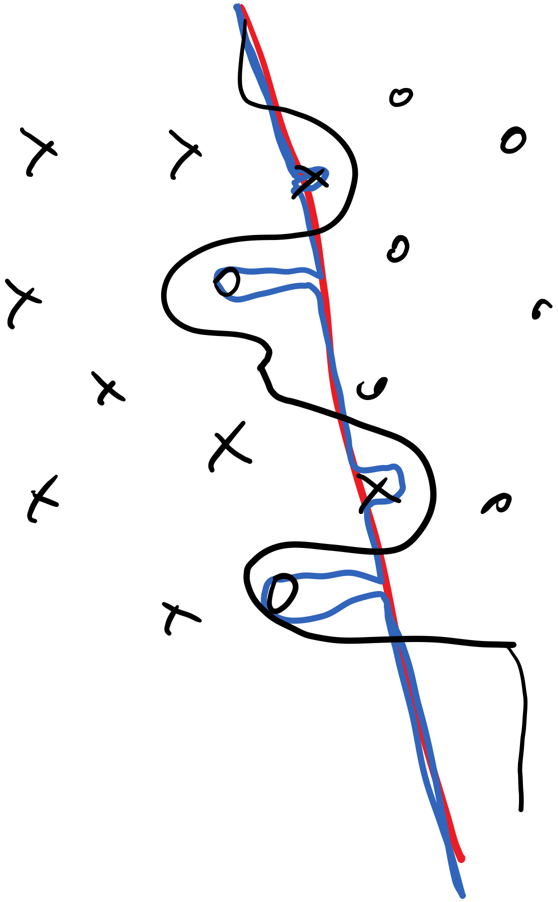
class labels

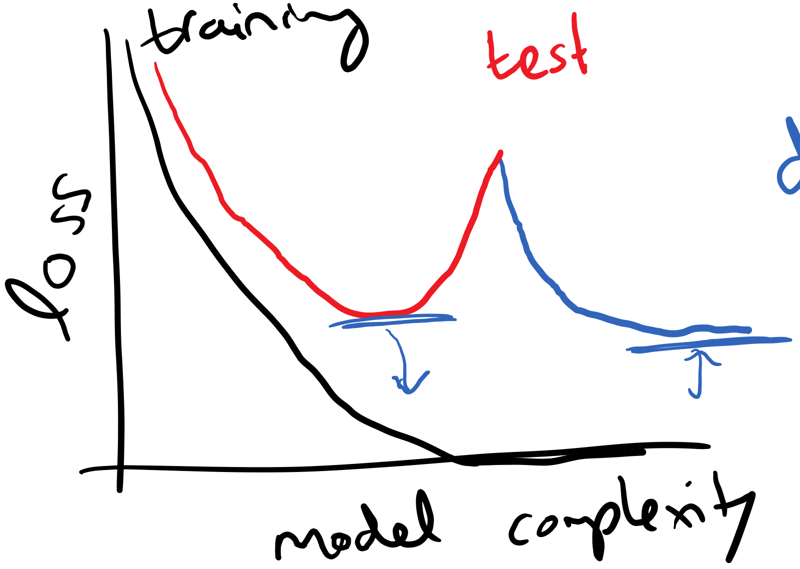
Confusion matrix



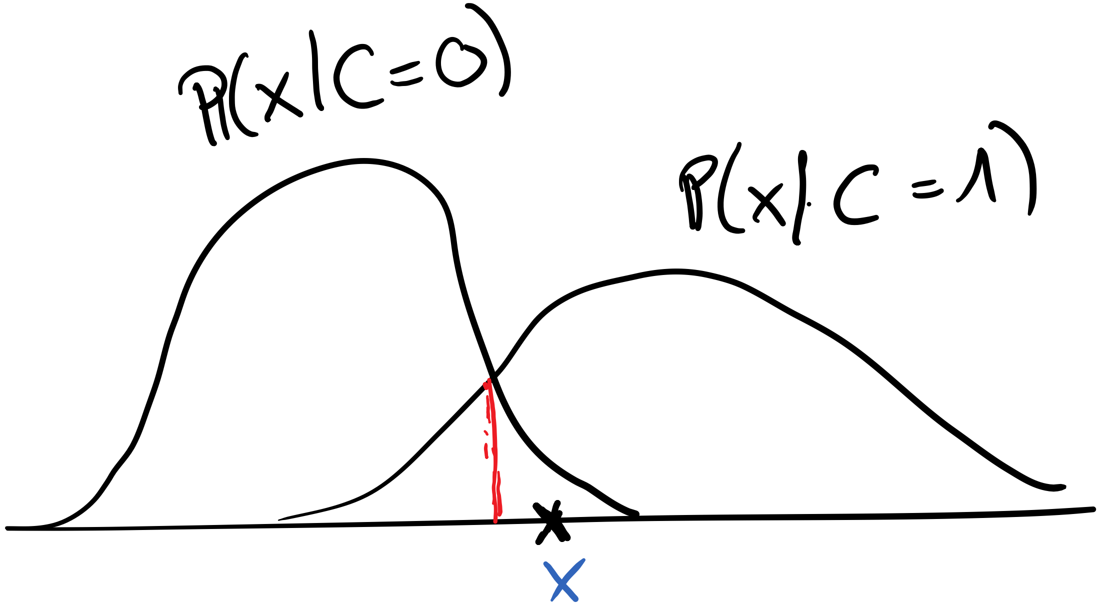
SVM







double descent



$P(C=0) >> P(C=1)$?

Best guess $x \in C=1$

$$\text{Estimate} \left.\begin{array}{l} P(x|C) \sim \mathcal{N}(\mu_c, \Sigma_c) \\ P(C) \end{array}\right\} \text{QDA}$$

$\mathbb{P}(C=0 \mid x)$ $\qquad$ $\mathbb{P}(C=1 \mid x)$

$[0,1]$ $\qquad$ $\beta_0 + \beta^T x \in (-\infty, \infty)$

$$\log \frac{\mathbb{P}(C=0 \mid x)}{\mathbb{P}(C=1 \mid x)} = \beta_0 + \beta^T x + \lambda \|\beta\|_1$$

$\Rightarrow \beta$ sparse

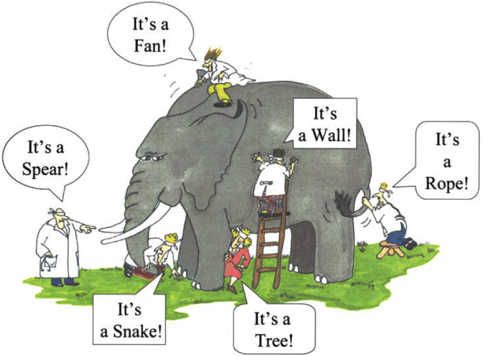
MITx: Statistics, Computation & Applications

Genomics and High-Dimensional Data Module

Lecture 2: Classification with Hig-Dimensional Data

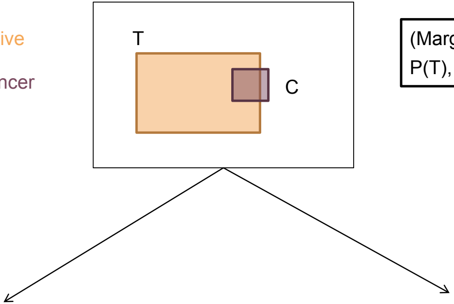
Classification with high-dimensional data

- Linear / Quadratic discriminant analysis
- Logistic regression
- Support Vector Machines



Refresher: Bayes rule

Sample space



T: Med. Test positive

C: Patient has cancer

(Marginal) Probability:
P(T), P(C)

New sample space:
People with cancer

P(T|C)
large

Conditional Probability:
P(T|C), P(C|T)

New sample space:
People with pos. test

P(C|T)
small

Bayes Theorem:

posterior

$$\boxed{P(C|T)} = \frac{\boxed{P(T|C)}\boxed{P(C)}}{P(T)}$$

prior

Class conditional probability

Using Bayes rule for classification

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \sim P(C)P(X|C)$$

Find some estimate

Prior / prevalence:
Fraction of samples
in that class

Assume:
$X|C \sim N(\mu_c, \Sigma_c)$

- Choose class $c \in \{1, \ldots, K\}$ such that $P(C = c \mid X)$ is maximal

Using Bayes rule for classification

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \sim P(C)P(X|C)$$

Find some estimate

Prior / prevalence:
Fraction of samples
in that class

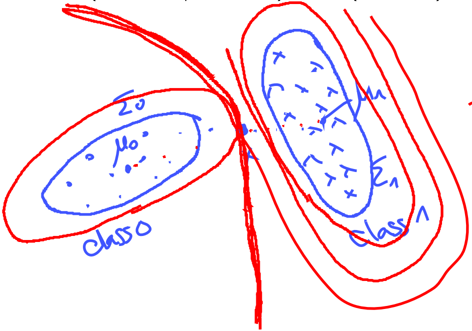
Assume:
$X|C \sim N(\mu_c, \Sigma_c)$

- Choose class $c \in \{1, \ldots, K\}$ such that $P(C = c \mid X)$ is maximal
- Special case: 2 classes 0/1
  - choose $c = 1$ if $P(C = 1 \mid X) > 0.5$
  - equivalently, choose $c = 1$ if posterior odds $P(C = 1 \mid X)/P(C = 0 \mid X) > 1$
- We need to estimate $P(C = c)$ an $P(X \mid C)$

estimate:
μ_c, Σ_c MLE

Quadratic discriminant analysis

- Assume $X \mid C = c \sim \mathcal{N}(\mu_c, \Sigma_c)$
- Estimate $P(C = c)$, μ_c, and Σ_c for each c (How?) MLE
- Choose class c such that
  $P(C = c \mid X = x) \propto P(C = c)P(X = x \mid C = c)$ is maximal



$\Rightarrow$ classify x as class 0.

Quadratic discriminant analysis

- Assume $X \mid C = c \sim \mathcal{N}(\mu_c, \Sigma_c)$
- Estimate $P(C = c)$, μ_c, and Σ_c for each c (How?)
- Choose class c such that
  $P(C = c \mid X = x) \propto P(C = c)P(X = x \mid C = c)$ is maximal
- Use the fact that maximizing $P(C = c \mid X = x)$ is equivalent to maximizing $\log(P(C = c \mid X = x))$
- Do the math:

$$\log(P(C = c \mid X = x)) \propto \log(P(C = c)) - \frac{1}{2}\log\det\Sigma_c - \frac{1}{2}(x - \mu_c)^T\Sigma_c^{-1}(x - \mu_c)$$

$$f_{\mu_c,\Sigma_c} \propto \frac{1}{\det(\Sigma_c)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_c)\Sigma^{-1}(x-\mu_c)\right)$$

Quadratic discriminant analysis

- Assume $X \mid C = c \sim \mathcal{N}(\mu_c, \Sigma_c)$
- Estimate $P(C = c)$, μ_c, and Σ_c for each c (How?)
- Choose class c such that
  $P(C = c \mid X = x) \propto P(C = c)P(X = x \mid C = c)$ is maximal
- Use the fact that maximizing $P(C = c \mid X = x)$ is equivalent to maximizing $\log(P(C = c \mid X = x))$
- Do the math:
  $$\log(P(C = c \mid X = x)) \propto \log(P(C = c)) - \frac{1}{2}\log\det\Sigma_c - \frac{1}{2}(x - \mu_c)^T\Sigma_c^{-1}(x - \mu_c)$$
- Decision boundaries are quadratic

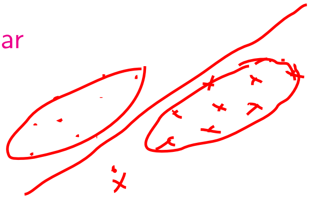
$\mathbb{P}(C=0 \mid X=x) = \mathbb{P}(C=1 \mid X=x)$

Linear discriminant analysis

- Assume same covariance matrix Σ in all classes, i.e. $X \mid C = c \sim \mathcal{N}(\mu_c, \Sigma)$
- Estimate $P(C = c)$, μ_c, and Σ for each c
- Choose class c such that $\log(P(C = c \mid X = x))$ is maximal
- Do the math:

$$\log(P(C = c \mid X = x)) \propto \log(P(C = c)) - \frac{1}{2}\mu_c^T \Sigma^{-1} \mu_c + x^T \Sigma^{-1} \mu_c$$

- Decision boundaries are linear



Linear discriminant analysis

- Assume same covariance matrix Σ in all classes, i.e. $X \mid C = c \sim \mathcal{N}(\mu_c, \Sigma)$
- Estimate $P(C = c)$, μ_c, and Σ for each c
- Choose class c such that $\log(P(C = c \mid X = x))$ is maximal
- Do the math:

$$\log(P(C = c \mid X = x)) \propto \log(P(C = c)) - \frac{1}{2}\mu_c^T \Sigma^{-1} \mu_c + x^T \Sigma^{-1} \mu_c$$

- Decision boundaries are linear

What happens if we assume $\Sigma = \sigma^2 I$?

$$X^T X = \Sigma = V \Lambda V^T$$

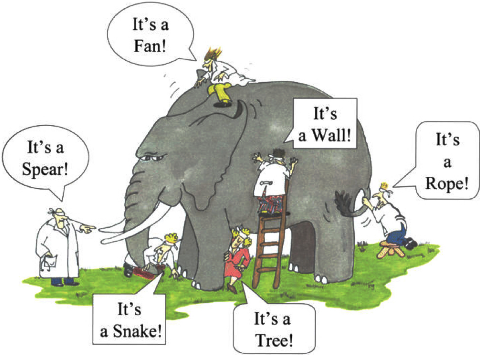
$$\Lambda^{-1/2} V^T (X^T X) V \Lambda^{-1/2} = \Lambda^{-1/2} \cancel{V^T} (\cancel{V} \Lambda \cancel{V^T}) \cancel{V} \Lambda^{-1/2}$$

$$= \mathrm{Id}$$

$$X \to X V \Lambda^{-1/2}$$

Classification with high-dimensional data

- Linear / Quadratic discriminant analysis
- Logistic regression
- Support Vector Machines



Reduced-rank LDA (a.k.a. Fisher's LDA)

How can we use LDA to find informative low-dimensional projection of the data?

Reduced-rank LDA (a.k.a. Fisher's LDA)

How can we use LDA to find informative low-dimensional projection of the data?

- idea: $\mu_1, \ldots, \mu_K \in \mathbb{R}^p$ lie in a linear subspace of dim $K-1$ (usually $p >> k$)
- if $K = 3$, then data can be projected into 2d
- if $K > 3$, combine LDA with PCA, i.e. perform PCA on class means
    - 1. LD is 1. PC of class means, 2. LD is 2. PC of class means, etc.

Reduced-rank LDA (a.k.a. Fisher's LDA)

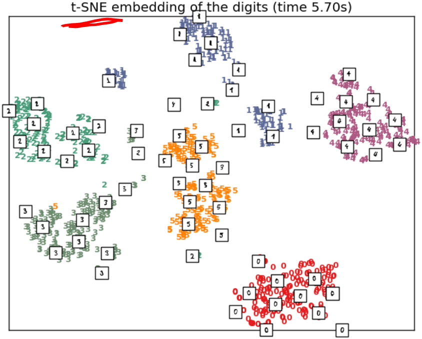
How can we use LDA to find informative low-dimensional projection of the data?

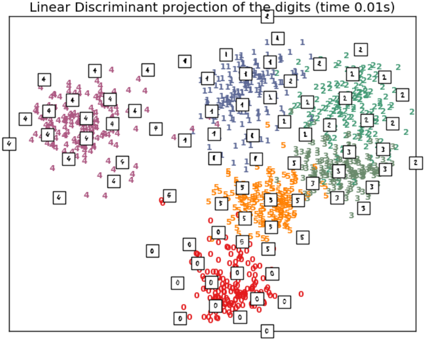
- idea: $\mu_1, \ldots, \mu_K \in \mathbb{R}^p$ lie in a linear subspace of dim $K-1$ (usually $p >> k$)
- if $K = 3$, then data can be projected into 2d
- if $K > 3$, combine LDA with PCA, i.e. perform PCA on class means
  - 1. LD is 1. PC of class means, 2. LD is 2. PC of class means, etc.

What is maximum number of LDs? $\min(p, \underline{k-1})$

Example: Digit recognition

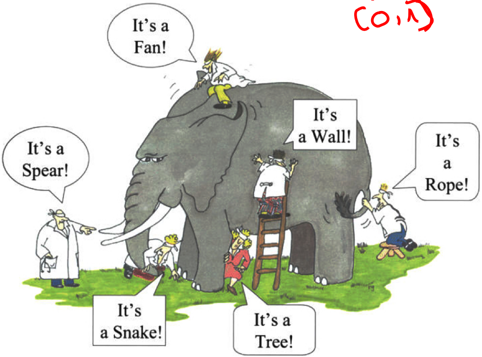
LDA can be used to find informative low-dimensional projection of the data by performing PCA on class means!





Classification with high-dimensional data

- Linear / Quadratic discriminant analysis
- Logistic regression
- Support Vector Machines



$C = \{0, 1\}$

$P(C=1 \mid x) = \beta_0 + \beta^T x$

$\in [0,1]$ $\qquad$ $\in (-\infty, \infty)$

$\log\left(\frac{p}{1-p}\right)$

$\in (-\infty, \infty)$

Logistic regression

- Assume we have two classes $C = \{0, 1\}$
- $Y \sim \mathrm{Bernoulli}(p)$, and we would like to model p
- What is wrong with assuming a linear model for p, i.e. $p = \beta_0 + \beta^T x$?
- In logistic regression we model: $\boxed{\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta^T x}$
- We need to estimate β_0 and β (How?)

$\binom{n}{k} p^k (1-p)^{n-k}$

- Then, we can solve for p, namely

$$\boxed{p = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)}}$$

$k \log(p) + (n-k) \log(1-p)$

- Choose $C = 1$ if $p > 0.5$

SVM

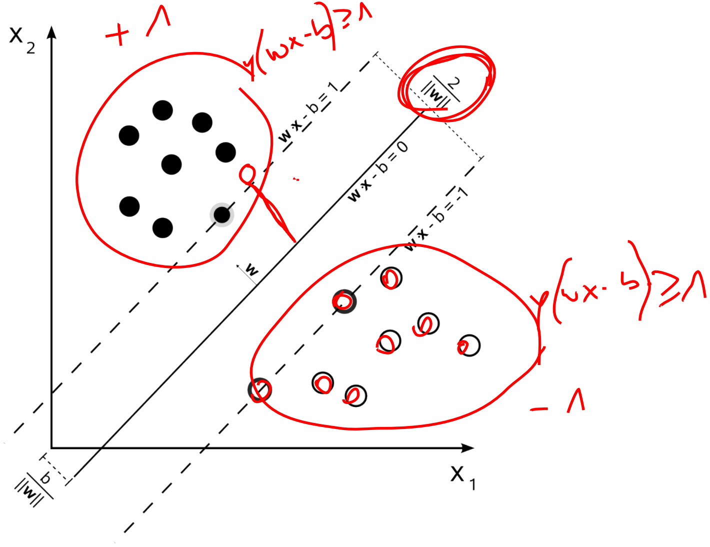


Figure from https://en.wikipedia.org/wiki/Support_vector_machine

SVM

- Given training data: $(x_1, y_1), \ldots, (x_n, y_n)$, with $x_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$
- If perfect classification is possible, determine hyperplane $(wx - b = 0)$ that maximizes distance to the nearest point x_i from each group:

$$\text{minimize } \|w\|_2 \quad \text{such that } \underline{y_i(wx_i - b) \geq 1} \quad \text{for all } i.$$

SVM

- Given training data: $(x_1, y_1), \ldots, (x_n, y_n)$, with $x_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$
- If perfect classification is possible, determine hyperplane $(wx - b = 0)$ that maximizes distance to the nearest point x_i from each group:

$$\text{minimize } \|w\|_2 \quad \text{such that } \; y_i(wx_i - b) \geq 1 \quad \text{for all } i.$$

- If perfect classification is not possible, determine hyperplane $(wx - b = 0)$ that maximizes distance to the nearest point x_i from each group and minimizes sum of classification errors φ_i:

$$\text{minimize } \|w\|_2 + \lambda \sum_{i=1}^{n} \varphi_i \; \text{ such that } \; \underline{y_i(wx_i - b) \geq 1 - \varphi_i} \quad \text{for all } i.$$

Summary: Classification

- **Bayesian decision theory**
  - Know probability distribution of the categories
  - Do not even need training data
  - Can design optimal classifier
- **Linear / Quadratic Discriminant Analysis (LDA / QDA)**
  - Shape of probability distribution of categories is known (Gaussian)
  - Need to estimate parameters of probability distribution
  - Need training data

  ← Logistic regression
- **Support Vector Machine (SVM)**
  - Non-parametric method, i.e., no probability distribution
  - Need to estimate parameters of discriminant function
  - Labeled data, need training data
- **Clustering**
  - Unlabeled data

Quality of classification

2 approaches:

- separate training data from test data
- cross validation, e.g. leave-one-out cross validation, where every sample is the test case once, the rest is the training data

Measures for prediction error:

- Build a confusion matrix

	Truth = 0	Truth = 1	Truth = 2
Estimate = 0	23	7	6
Estimate = 1	3	27	4
Estimate = 2	3	1	26

- Error rate = 1 - sum(diagonal entries) / (total number of samples)

References

Chapter 4 in

- T. Hastie, R. Tibshirani, & J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

I. Goodfellow, Y. Bengio, & A. Courville. *Deep Learning*. MIT Press, 2016.

M. Belkin, D. Hsu, S. Ma, & S. Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *PNAS* 116, 2019.