Supervised learning: Classification

Marta Belchior Lopes

NOVAMATH
CENTER FOR MATHEMATICS
+ APPLICATIONS

NOVA
NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

Where I am

UNIVERSIDADE NOVA
DE LISBOA

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

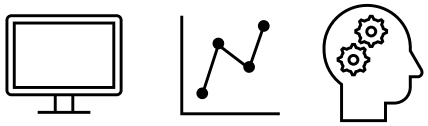
Where I am

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

NOVAMATH CENTER FOR MATHEMATICS + APPLICATIONS

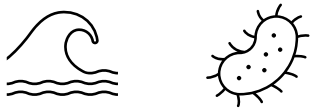
unidemi NOVA | RESEARCH & DEVELOPMENT UNIT FOR MECHANICAL AND INDUSTRIAL ENGINEERING

Biostatistics, Machine Learning and Bioinformatics



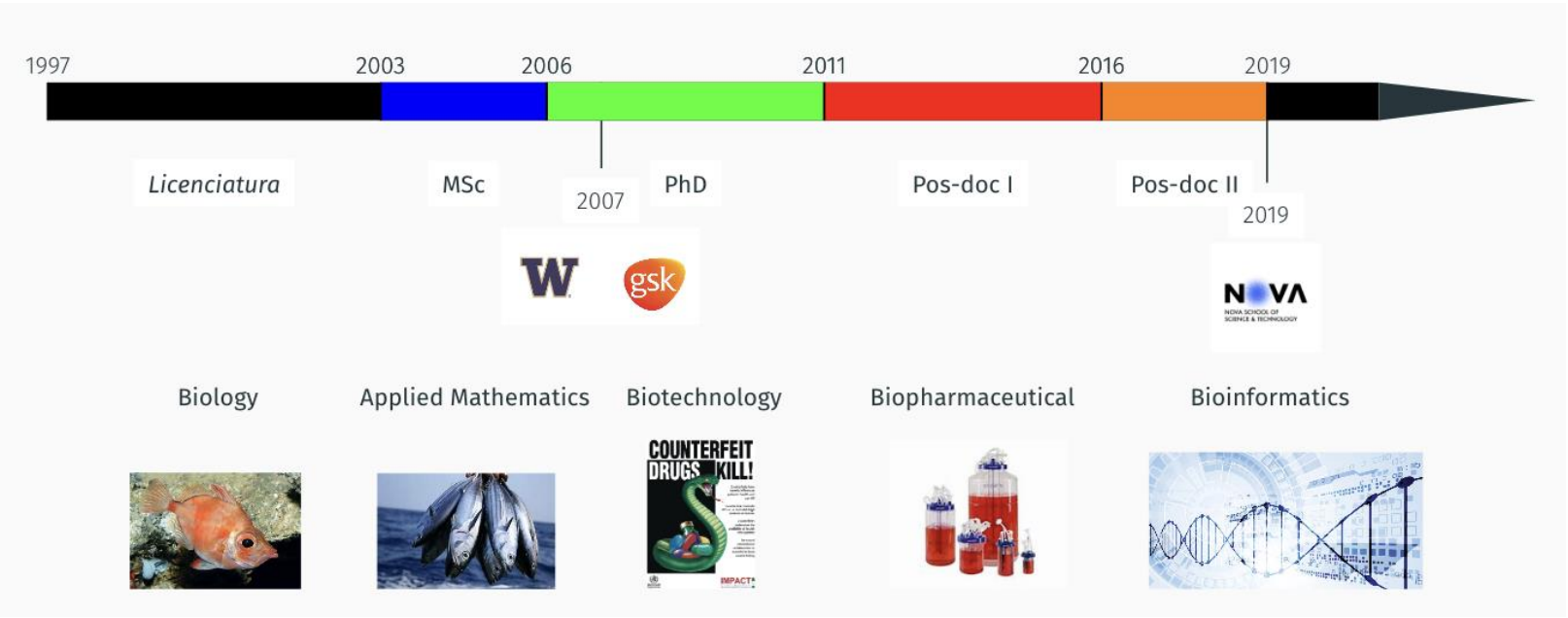
- Precision Medicine

- Environmental



- Ecology

My academic and research timeline



Generalized Linear Models

Linear regression

Multiple linear regression

- Numerical response
- Numerical and categorical predictors

Generalized Linear Models

Linear regression

Multiple linear regression

- Numerical response
- Numerical and categorical predictors

Other type of response?
(e.g., categorical, count data)

Generalized Linear Models

Linear regression

Multiple linear regression

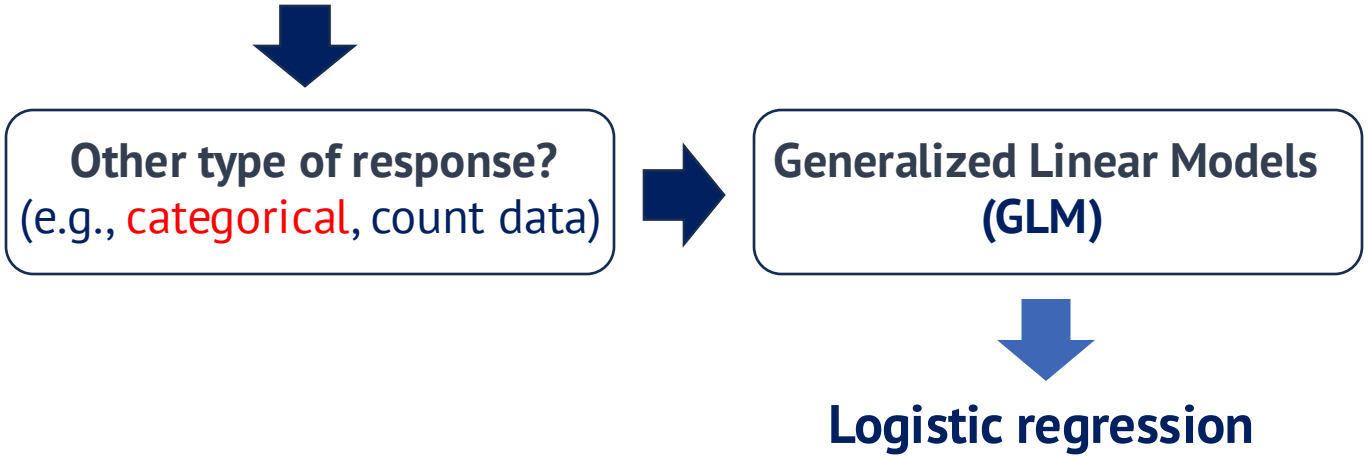
- Numerical response
- Numerical and categorical predictors

Generalized Linear Models

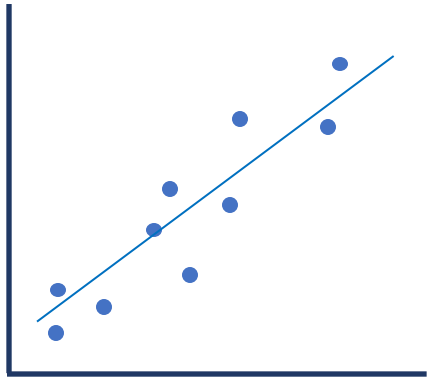
Linear regression

Multiple linear regression

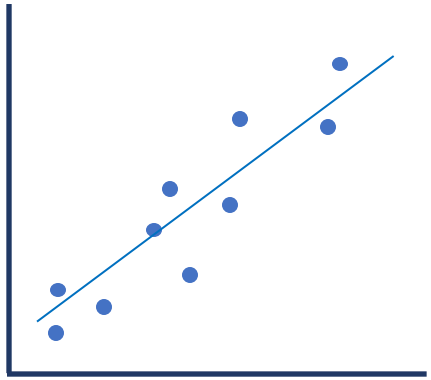
- Numerical response
- Numerical and categorical predictors

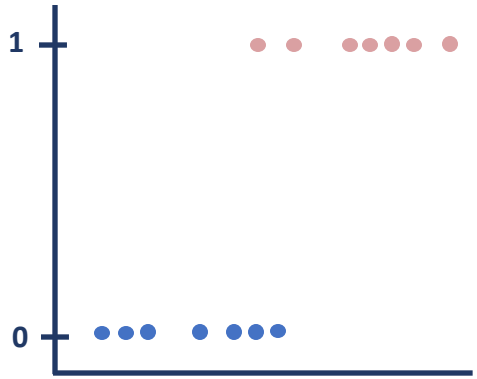


Logistic regression

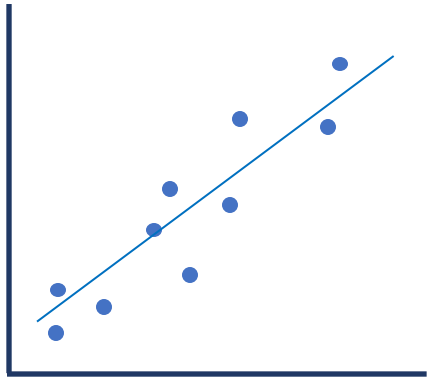


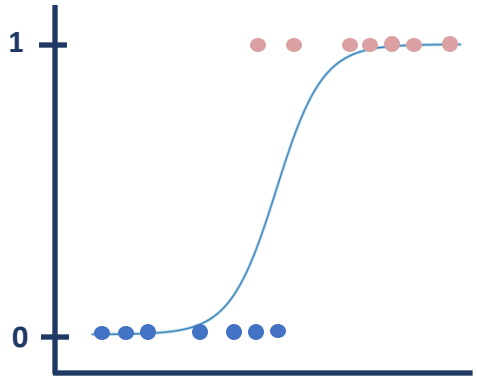
Logistic regression



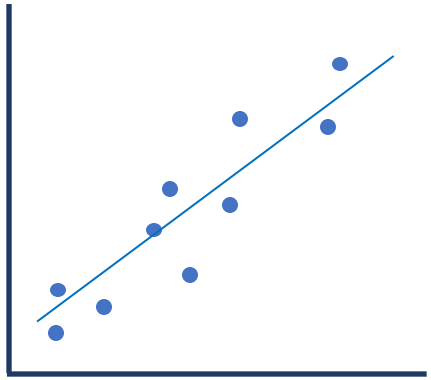


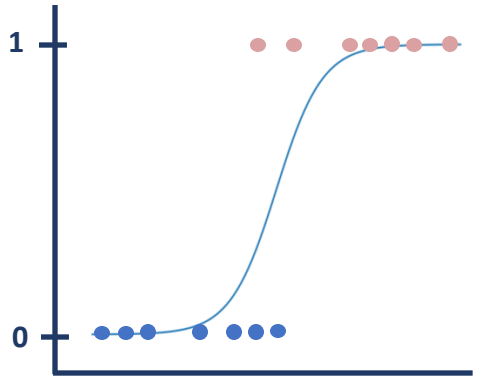
Logistic regression





Logistic regression





GLM uses a **link** function $g(.)$

Connects the linear predictor to the expected value of the response variable

Logistic regression

The **logit** function

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

p is the probability of success (event occurrence)

$1 - p$ is the probability of failure (non-event occurrence)

β_0 is the intercept term

$X_1, X_2, \ldots, X_p$ are the predictor variables

Logistic regression

The **logit** function

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

p is the probability of success (event occurrence)

$1 - p$ is the probability of failure (non-event occurrence)

β_0 is the intercept term

$X_1, X_2, \ldots, X_p$ are the predictor variables

The **logistic** function

$$p = P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}}$$

$P(Y = 1)$ is the probability that the outcome variable Y is equal to 1 (event)

Logistic regression

The **logit** function

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

p is the probability of success (event occurrence)

$1 - p$ is the probability of failure (non-event occurrence)

β_0 is the intercept term

$X_1, X_2, \ldots, X_p$ are the predictor variables

The **logistic** function

$$p = P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)}}$$

$P(Y = 1)$ is the probability that the outcome variable Y is equal to 1 (event)

$$l(\beta_0, \beta) = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log p_i + (1 - y_i) \log(1 - p_i) \right]$$

Logistic regression

BIOLOGY, ECOLOGY & HEALTH

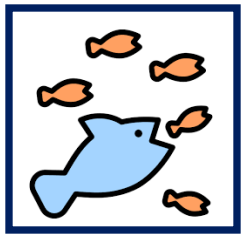
Presence or **absence** of a species in a given *habitat* based on environmental variables (e.g., temperature, humidity, vegetation cover)

flaticon.com

Logistic regression

BIOLOGY, ECOLOGY & HEALTH

Presence or **absence** of a species in a given *habitat* based on environmental variables (e.g., temperature, humidity, vegetation cover)

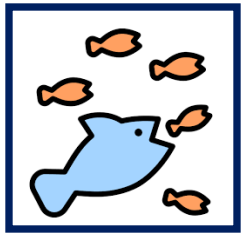


Predator-prey relationships, to predict predatory events based on variables such as predator abundance, environmental conditions, prey density

Logistic regression

BIOLOGY, ECOLOGY & HEALTH

Presence or **absence** of a species in a given *habitat* based on environmental variables (e.g., temperature, humidity, vegetation cover)



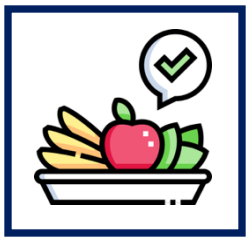
Predator-prey relationships, to predict predatory events based on variables such as predator abundance, environmental conditions, prey density

Contamination events, based on biological and environmental variables (pollution, contaminant, climate)

flaticon.com

Logistic regression

BIOLOGY, ECOLOGY & HEALTH



Food safety, based variables measured by analytical testing in the laboratory

flaticon.com

Logistic regression

BIOLOGY, ECOLOGY & HEALTH

Food safety, based variables measured by analytical testing in the laboratory

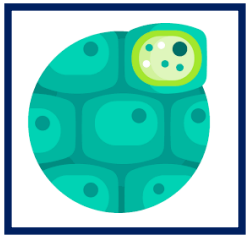
Disease presence-absence in a given population based on e.g., clinical, demographic and environmental variables

Logistic regression

BIOLOGY, ECOLOGY & HEALTH

Food safety, based variables measured by analytical testing in the laboratory

Disease presence-absence in a given population based on e.g., clinical, demographic and environmental variables



Disease/cell subtype, based on molecular variables (e.g., gene expression)

flaticon.com

Logistic regression

Model training and testing

- Data split into training and test sets

Logistic regression

Model training and testing

- Data split into training and test sets
- Train the logistic regression model

Logistic regression

Model training and testing

- Data split into training and test sets
- Train the logistic regression model
- Make predictions on the test set

Logistic regression

Model training and testing

- Data split into training and test sets
- Train the logistic regression model
- Make predictions on the test set
- Evaluate the model

Logistic regression

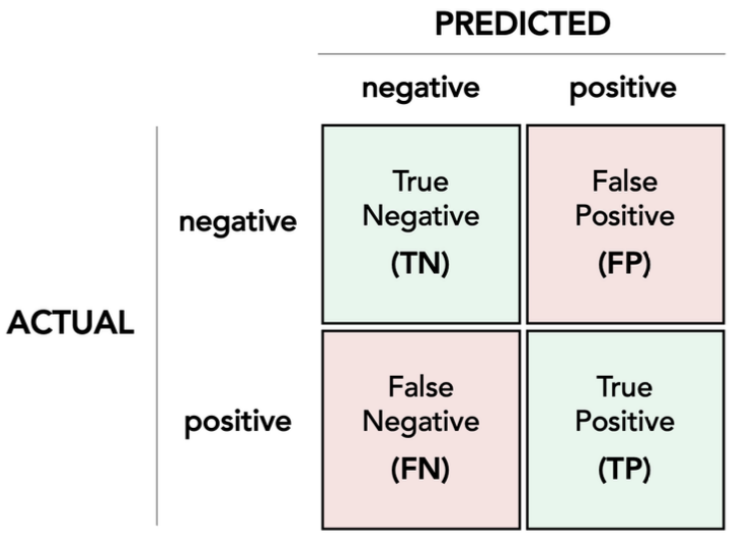
Model evaluation

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN} \qquad \text{FPR} = \frac{FP}{FP + TN}$$

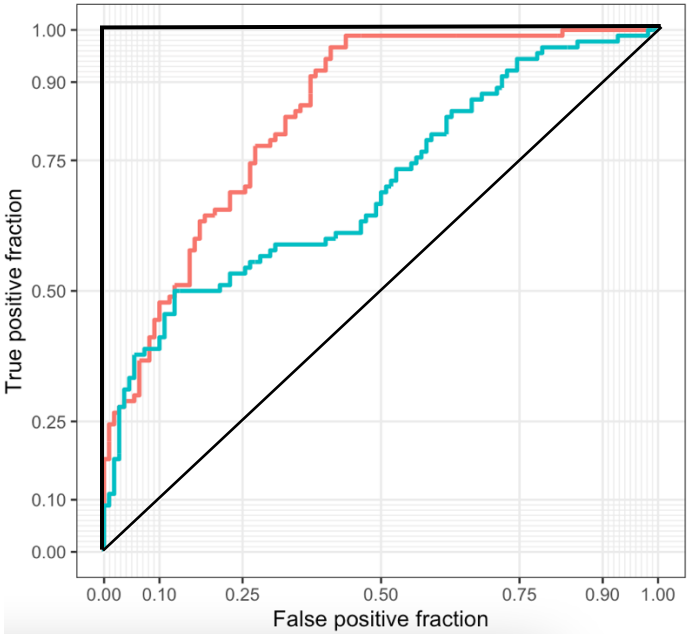
$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP + FP + FN}$$



Logistic regression

Model evaluation

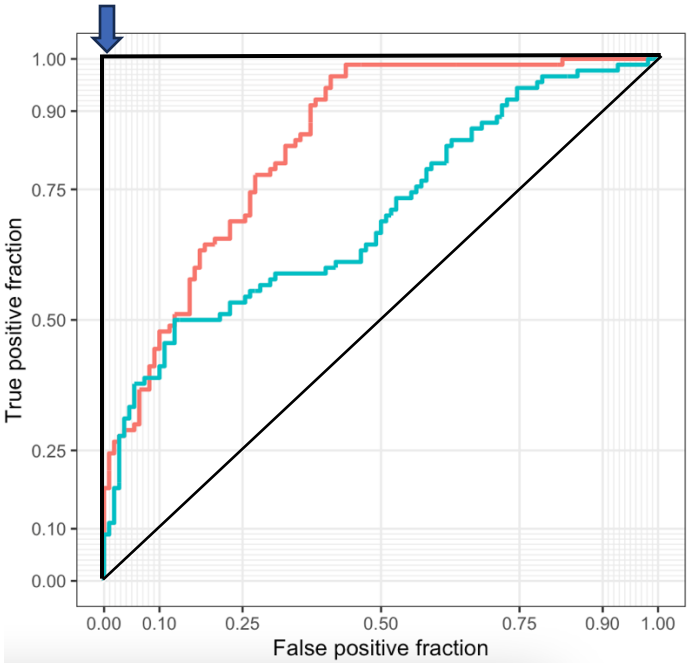
- Receiver Operating Characteristic (**ROC**) curve
- Area under the ROC curve (**AUC**)



Logistic regression

Model evaluation

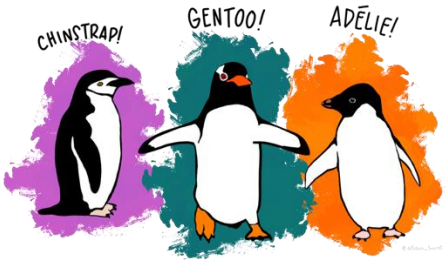
- Receiver Operating Characteristic (**ROC**) curve
- Area under the ROC curve (**AUC**)



Logistic regression

Penguin species (Palmer Archipelago)

Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0.

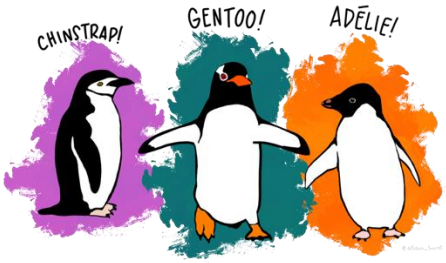


- bill_length_mm
- flipper_length_mm
- body_mass_g

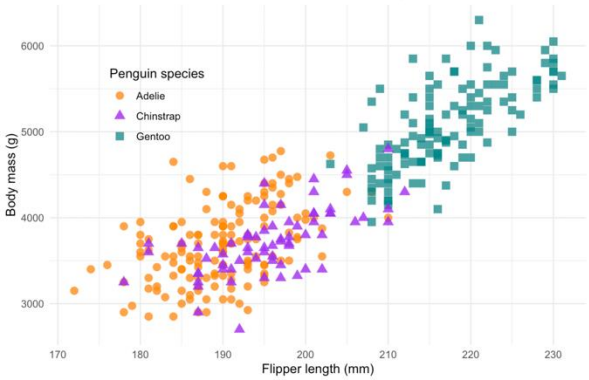
Logistic regression

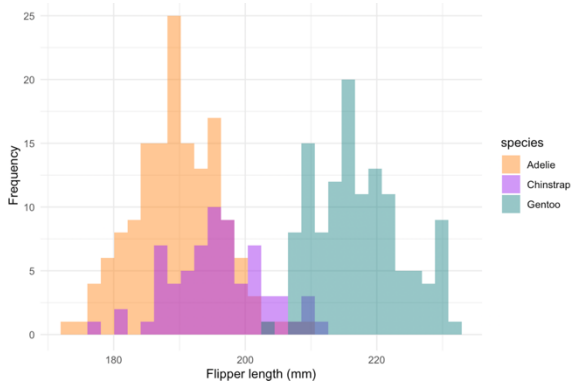
Penguin species (Palmer Archipelago)

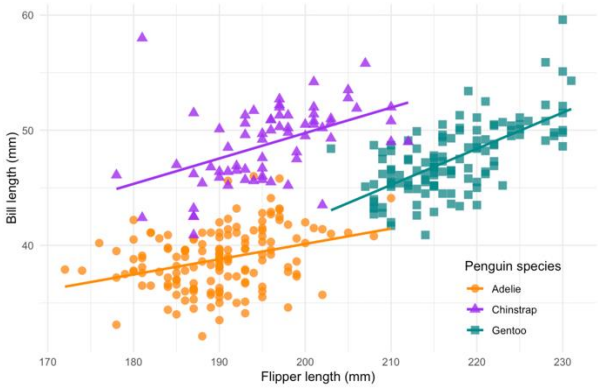
Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0.



- bill_length_mm
- flipper_length_mm
- body_mass_g







Logistic regression

Penguin species (Palmer Archipelago)

- Create a binary outcome variable “IsGentoo”
- Partition the data into a training and a test set
- Build a binary logistic regression with multiple predictor variables
- Make predictions on the test set
- Evaluate the model using the confusion matrix and AUC

Logistic regression

Penguin species (Palmer Archipelago)

```r
# Install packages and load the penguins dataset
install.packages("caret")
install.packages("pROC")
install.packages("palmerpenguins")

library("palmerpenguins")
data("penguins")
dim(penguins)
head(penguins,10)
penguins <- na.omit(penguins)  # remove rows with missing values

# Create a binary outcome variable, TRUE if species is Gentoo
penguins$IsGentoo <- ifelse(penguins$species == "Gentoo", 1, 0)

# Partition the data into a training set and a test set, preserving class distribution
set.seed(123)  # For reproducibility
library("caret")
partition <- createDataPartition(penguins$IsGentoo, p = 0.7, list = FALSE)
train_data <- penguins[partition, ]
test_data <- penguins[-partition, ]

# Create a binary logistic regression model with multiple predictor variables
model <- glm(IsGentoo ~ bill_length_mm + flipper_length_mm + body_mass_g,
             data = train_data, family = binomial)

# Make predictions on the test set
test_predictions <- predict(model, newdata = test_data, type = "response")
```

Logistic regression

Penguin species (Palmer Archipelago)

```r
## Evaluate the model

# Convert probabilities to class labels (threshold = 0.5)
test_labels <- ifelse(test_predictions > 0.5, TRUE, FALSE)

confusion_matrix <- table(Predicted = test_labels,
                          Actual = test_data$IsGentoo)
print(confusion_matrix)

# Compute the AUC (Area Under the ROC Curve)
library(pROC)
roc_obj <- roc(test_data$IsGentoo, test_predictions)  # probs from glm
auc_value <- auc(roc_obj)
auc_value

plot(roc_obj, main = "ROC Curve")
```

Logistic regression

Penguin species (Palmer Archipelago)

```
> summary(model)


  Call:
  glm(formula = IsGentoo ~ bill_length_mm + flipper_length_mm +
      body_mass_g, family = binomial, data = train_data)

  Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
  (Intercept)       -1.230e+02  3.201e+01  -3.844 0.000121 ***
  bill_length_mm    -3.340e-01  1.693e-01  -1.973 0.048492 *
  flipper_length_mm  5.631e-01  1.610e-01   3.498 0.000469 ***
  body_mass_g        4.954e-03  2.086e-03   2.376 0.017518 *
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

  (Dispersion parameter for binomial family taken to be 1)

      Null deviance: 308.835  on 233  degrees of freedom
  Residual deviance:  25.393  on 230  degrees of freedom
  AIC: 33.393
```

Logistic regression

Penguin species (Palmer Archipelago)

```
> summary(model)

Call:
glm(formula = IsGentoo ~ bill_length_mm + flipper_length_mm +
    body_mass_g, family = binomial, data = train_data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.230e+02  3.201e+01  -3.844 0.000121 ***
bill_length_mm -3.340e-01  1.693e-01  -1.973 0.048492 *
flipper_length_mm 5.631e-01  1.610e-01   3.498 0.000469 ***
body_mass_g     4.954e-03  2.086e-03   2.376 0.017518 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 308.835  on 233  degrees of freedom
Residual deviance:  25.393  on 230  degrees of freedom
AIC: 33.393
```

$$\log\left(\frac{p}{1-p}\right) = -1.230\text{e+}02 - 3.340\text{e-}01 \times \text{bill\_length\_mm} + 5.631\text{e-}01 \times \text{flipper\_length\_mm} + 4.954\text{e-}03 \times \text{body\_mass\_mm}$$

Logistic regression

Penguin species (Palmer Archipelago)

```
> summary(model)

Call:
glm(formula = IsGentoo ~ bill_length_mm + flipper_length_mm +
    body_mass_g, family = binomial, data = train_data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.230e+02  3.201e+01  -3.844 0.000121 ***
bill_length_mm -3.340e-01  1.693e-01  -1.973 0.048492 *
flipper_length_mm 5.631e-01  1.610e-01   3.498 0.000469 ***
body_mass_g     4.954e-03  2.086e-03   2.376 0.017518 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 308.835  on 233  degrees of freedom
Residual deviance:  25.393  on 230  degrees of freedom
AIC: 33.393
```

$$\log\left(\frac{p}{1-p}\right) = -1.230\text{e+}02 - 3.340\text{e-}01 \times \text{bill\_length\_mm} + 5.631\text{e-}01 \times \text{flipper\_length\_mm} + 4.954\text{e-}03 \times \text{body\_mass\_mm}$$

bill_length = 45 mm
flipper_length = 220 mm
body_mass = 5500 g

$$\log\left(\frac{p}{1-p}\right) = 13.1 \Rightarrow p = 0.99$$

Logistic regression

Penguin species (Palmer Archipelago)

```
> summary(model)

Call:
glm(formula = IsGentoo ~ bill_length_mm + flipper_length_mm +
    body_mass_g, family = binomial, data = train_data)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.230e+02  3.201e+01  -3.844 0.000121 ***
bill_length_mm -3.340e-01  1.693e-01  -1.973 0.048492 *
flipper_length_mm 5.631e-01  1.610e-01   3.498 0.000469 ***
body_mass_g     4.954e-03  2.086e-03   2.376 0.017518 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 308.835  on 233  degrees of freedom
Residual deviance:  25.393  on 230  degrees of freedom
AIC: 33.393
```

$$\log\left(\frac{p}{1-p}\right) = -1.230\text{e+}02 - 3.340\text{e-}01 \times \text{bill\_length\_mm} + 5.631\text{e-}01 \times \text{flipper\_length\_mm} + 4.954\text{e-}03 \times \text{body\_mass\_mm}$$

bill_length = 45 mm
flipper_length = 220 mm
body_mass = 5500 g

$$\log\left(\frac{p}{1-p}\right) = \mathbf{13.1} \Rightarrow p = \mathbf{0.99}$$

bill_length = 45 mm
flipper_length = 195 mm
body_mass = 5500 g

$$\log\left(\frac{p}{1-p}\right) = \mathbf{-0.98} \Rightarrow p = \mathbf{0.27}$$

Summary

- **Logistic regression** for binary outcome variables
- **Examples** of application in Biology, Ecology and Health
- **Mathematical** formulation and interpretation
- **R code** example

Fundamentals of
Biodata Analysis with R

December 10-13, 2025

FNS, University of Tirana, Albania

Dimensionality reduction and feature selection

Marta Belchior Lopes

NOVAMATH
CENTER FOR MATHEMATICS
+ APPLICATIONS

NOVA
NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

Dimensionality reduction

Why reducing data dimension?

- Not all features are important (e.g., less impact on the outcome or noise features)
- High number of features increase model complexity ➡ need to simplify
- Measuring can be expensive and time-consuming

Dimensionality reduction

Why reducing data dimension?

- Not all features are important (e.g., less impact on the outcome or noise features)
- High number of features increase model complexity ➡ need to simplify
- Measuring can be expensive and time-consuming

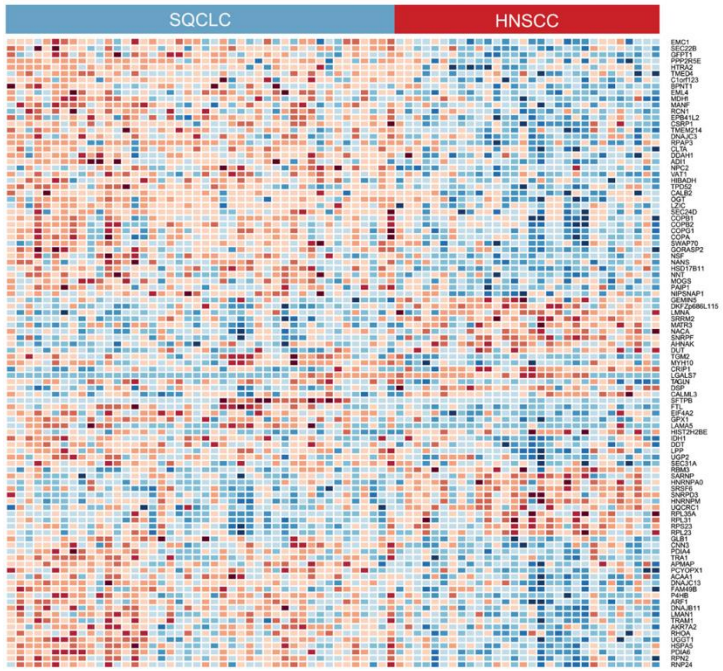
FEATURE SELECTION

➡ **Improve model performance, interpretability** and **efficiency**

Feature selection

Examples

- **Genomics and Proteomics** in disease modeling

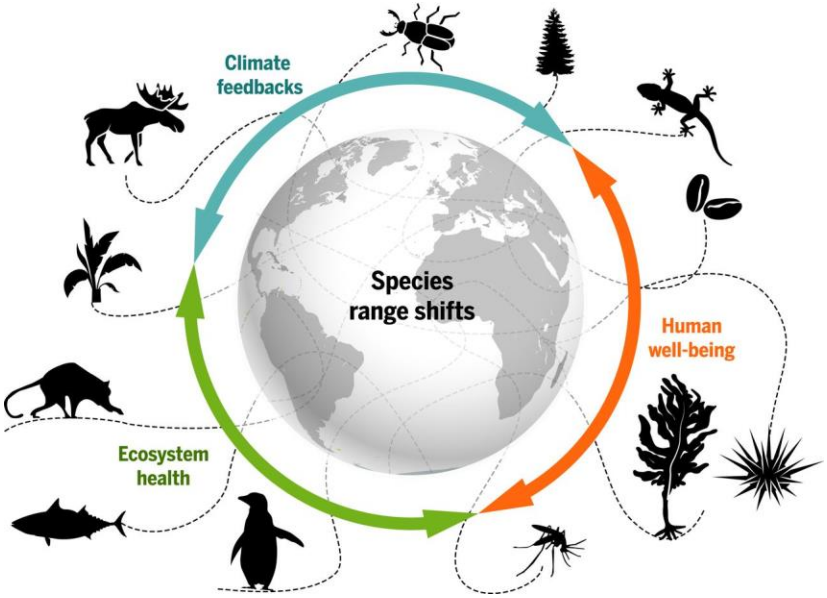


EMBO Mol Med (2018)10:e8428

Feature selection

Examples

- **Genomics and Proteomics** in disease modeling
- **Ecological** studies
  - Species distribution
  - Biodiversity

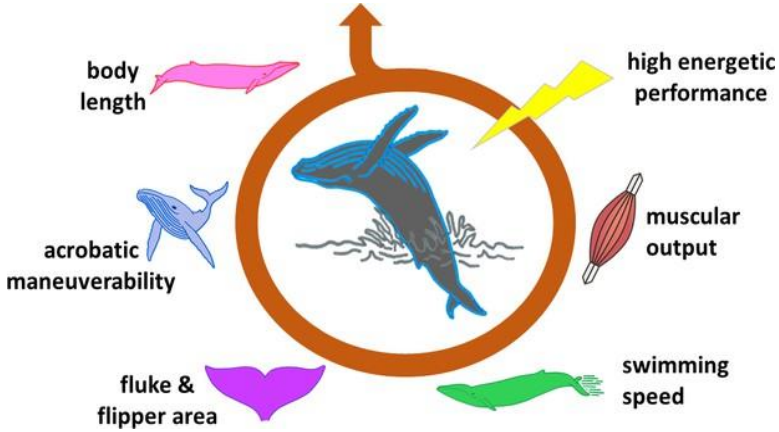


Science **355**, eaai9214 (2017)

Feature selection

Examples

- **Genomics and Proteomics** in disease modeling
- **Ecological** studies
  - Species distribution
  - Biodiversity
  - Behaviour ecology

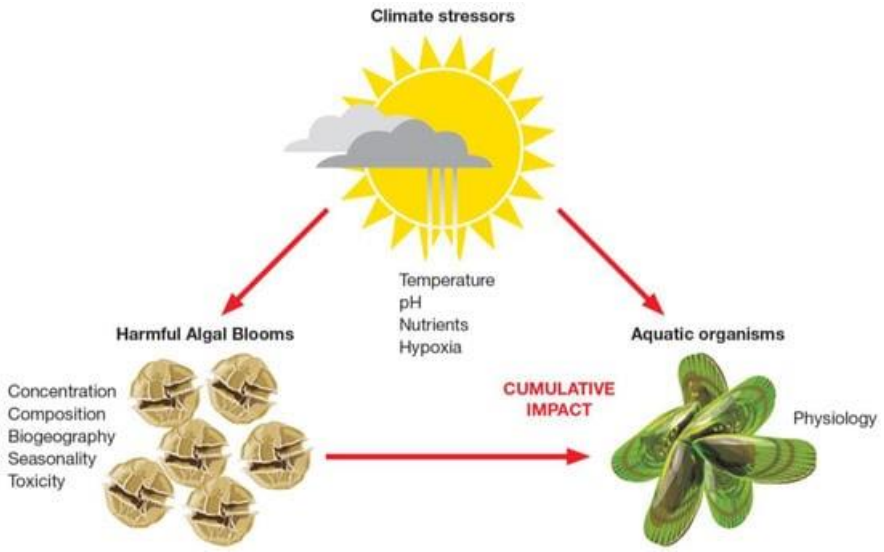


eLife **9**: e55722 (2022)

Feature selection

Examples

- **Genomics and Proteomics** in disease modeling
- **Ecological** studies
- **Toxicology** studies



Toxins **2022**, *14*, 341

Feature selection

Strategies

- **Filter** methods: independent of the learning algorithm used for model building (e.g., statistical properties of the variables)
- **Wrapper** methods: evaluate feature subsets using a learning algorithm and select features based on the performance of the algorithm
- **Embedded** methods: combine feature selection and model training, selecting features as part of the model-building process

Feature selection

Filter methods

- Independence on the learning algorithm, and used as a **preprocessing** step
- Computationally **efficient**
- Features assessed based on **statistical** measures

Feature selection

Filter methods

- Univariate (unsupervised and supervised)
    - Variance
    - Correlation
    - Mutual information
    - Chi-square test
    - Analysis of variance (ANOVA)

Feature selection

Filter methods

- Univariate (unsupervised and supervised)
    - Variance
    - Correlation
    - Mutual information
    - Chi-square test
    - Analysis of variance (ANOVA)
- Multivariate
    - Correlation-based

Feature selection

Wrapper methods

- **Dependence** on the learning algorithm
- **Iterative** process
- **Model performance** metrics as criterion (e.g., accuracy)
- Computationally **expensive** (inadequate for high-dimenisonal data)

Feature selection

Wrapper methods

- **Dependence** on the learning algorithm
- **Iterative** process
- **Model performance** metrics as criterion (e.g., accuracy)
- Computationally **expensive** (inadequate for high-dimenisonal data)
    - Forward selection
    - Backward elimination
    - Recursive feature elimination

Feature selection

Embedded methods

- Combined **feature selection** and **model training**
- More **efficient** than wrapper methods
    - L1 regularization (LASSO)
    - Tree-based methods

Feature selection

Embedded methods

- Combined **feature selection** and **model training**
- More **efficient** than wrapper methods
    - L1 regularization (LASSO)
    - Tree-based methods

Sparse logistic regression

$$l(\beta_0, \boldsymbol{\beta}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \pi_i + (1-y_i)\log(1-\pi_i)\right] + \lambda F_\alpha(\boldsymbol{\beta})$$

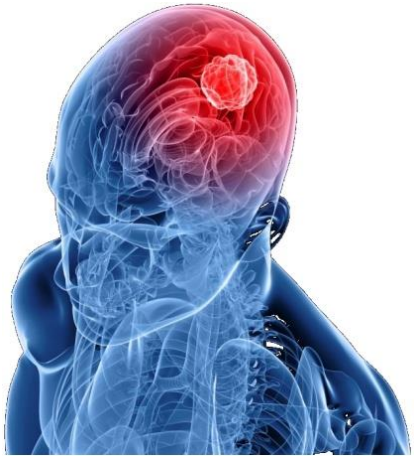
$$F_\alpha(\boldsymbol{\beta}) = \left(\alpha \left\|\boldsymbol{\beta}\right\|_1 + \frac{1-\alpha}{2}\left\|\boldsymbol{\beta}\right\|_2^2\right), \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$$

Elastic net & LASSO

Feature selection

Single-cell gene expression data from gliomas

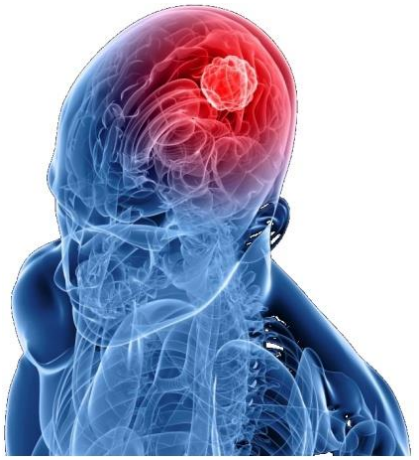
Gliomas – the most common brain tumors



Feature selection

Single-cell gene expression data from gliomas

Gliomas – the most common brain tumors



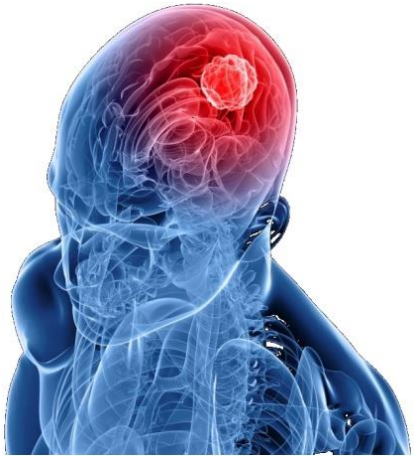
Intertumoral heterogeneity

- Astrocytoma
- Oligodendroglioma
- Glioblastoma

Feature selection

Single-cell gene expression data from gliomas

Gliomas – the most common brain tumors



Intertumoral heterogeneity

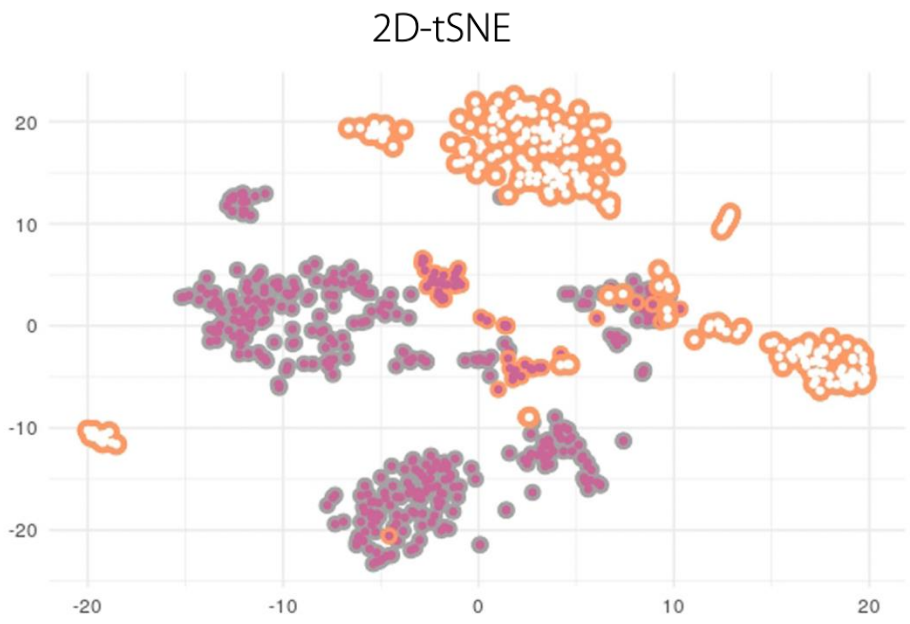
- Astrocytoma
- Oligodendroglioma
- Glioblastoma

Intratumoral heterogeneity

Therapy failure and tumor relapse

Feature selection

Glioblastoma single-cell gene expression data



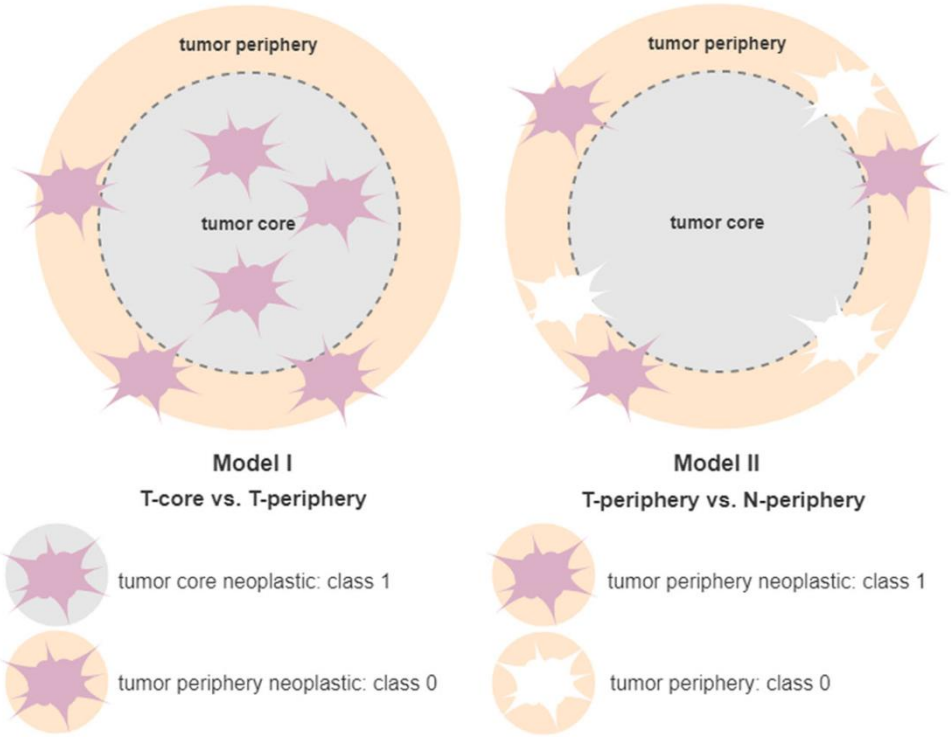
- Single-cell RNA-seq data
- Four primary glioblastoma patients
- **3,589 cells** and **23,368 genes**

tumor core neoplastic
tumor periphery neoplastic
tumor periphery normal

Lopes and Vinga *BMC Bioinformatics* (2020) 21:59
https://doi.org/10.1186/s12859-020-3390-4

Feature selection

Glioblastoma single-cell gene expression data



- **Sparse** logistic regression
- **Elastic net** penalty

Lopes and Vinga *BMC Bioinformatics* (2020) 21:59
https://doi.org/10.1186/s12859-020-3390-4

Feature selection

Glioblastoma single-cell gene expression data

Model performance

Classes	Vars	Miscl		AUC	
		Train	Test	Train	Test
I - T-core vs. T-periphery	83	10	7	0.97	0.94
II - T-periphery vs. N-periphery	85	3	4	0.99	0.96

Lopes and Vinga *BMC Bioinformatics* (2020) 21:59
https://doi.org/10.1186/s12859-020-3390-4

Feature selection

Glioblastoma single-cell gene expression data

Model performance

Classes	Vars	Miscl		AUC	
		Train	Test	Train	Test
I - T-core vs. T-periphery	83	10	7	0.97	0.94
II - T-periphery vs. N-periphery	85	3	4	0.99	0.96

Genes selected

Model I - T-core vs. T-periphery				
∗***ATP1A2***	***CLDN10***	***ECHDC2***	***FGFR3***	***GRM3***
HERC6	***HIF3A***	***HSPB8***	***NPL***	***PCSK1N***
PPM1K	∗***PRODH***	***SCG3***	***SPARCL1***	***TMSB10***
Model II - T-periphery vs. N-periphery				
ADAMTS3	***ADAMTSL1***	∗***ANXA1***	***COL28A1***	***CRNDE***
∗***EGFR***	***EMP1***	***F2R***	***GNG5***	***HES6***
HLA-A	***HOXB3***	***HSPB6***	∗***HTRA1***	***ID3***
∗***IFI44L***	***IGFBP2***	***IQCE***	***LINC00475***	***MGLL***
PSPH	∗***PTGDS***	***SEC61G***	***SPOCK1***	***VIM***

Lopes and Vinga *BMC Bioinformatics* (2020) 21:59
https://doi.org/10.1186/s12859-020-3390-4

Summary

- **Need** for data dimensionality reduction
- **Methods** for feature selection
- **Embedded** feature selection
- **Example** of application in Biomedicine