analiza danych

1. Wprowadzenie
2. Regresja liniowa – przypomnienie i najprostszy klasyfikator - regresja logistyczna
3. **Metody wyboru zmiennych oparte na równaniu regresji**
4. Metody preselekcji zmiennych
5. Transformacje – embedding ①
6. Transformacje – embedding ②
7. Prezentacja projektów

Plan wykładu

1. Literatura ①
2. Tradycyjne metody wyboru zmiennych
   - eliminacja wsteczna – backward elimination
   - selekcja wprzód – forward selection
   - selekcja krokowa – stepwise selection
   - walidacja krzyżowa – cross validation
3. Regresja z regularyzacją – Penalized regression
   - Lasso
   - Ridge
   - adaptive Lasso
   - elastic NET

Plan wykładu

4. IC
   - AIC
   - BIC
   - Modyfikacje
5. Kryteria dla modeli zagnieżdżonych
   - Test ilorazu prawdopodobieństwa – Likelihood Ratio Test
   - Deviance
6. Literatura ②

1. **Literatura ①**
2. Tradycyjne metody wyboru zmiennych
3. Regresja z regularyzacją – Penalized regression
4. IC
5. Kryteria dla modeli zagnieżdżonych
6. Literatura ②

Literatura

Supervised Rank Aggregation (SRA): A Novel Rank Aggregation Approach for Ensemble-based Feature Selection

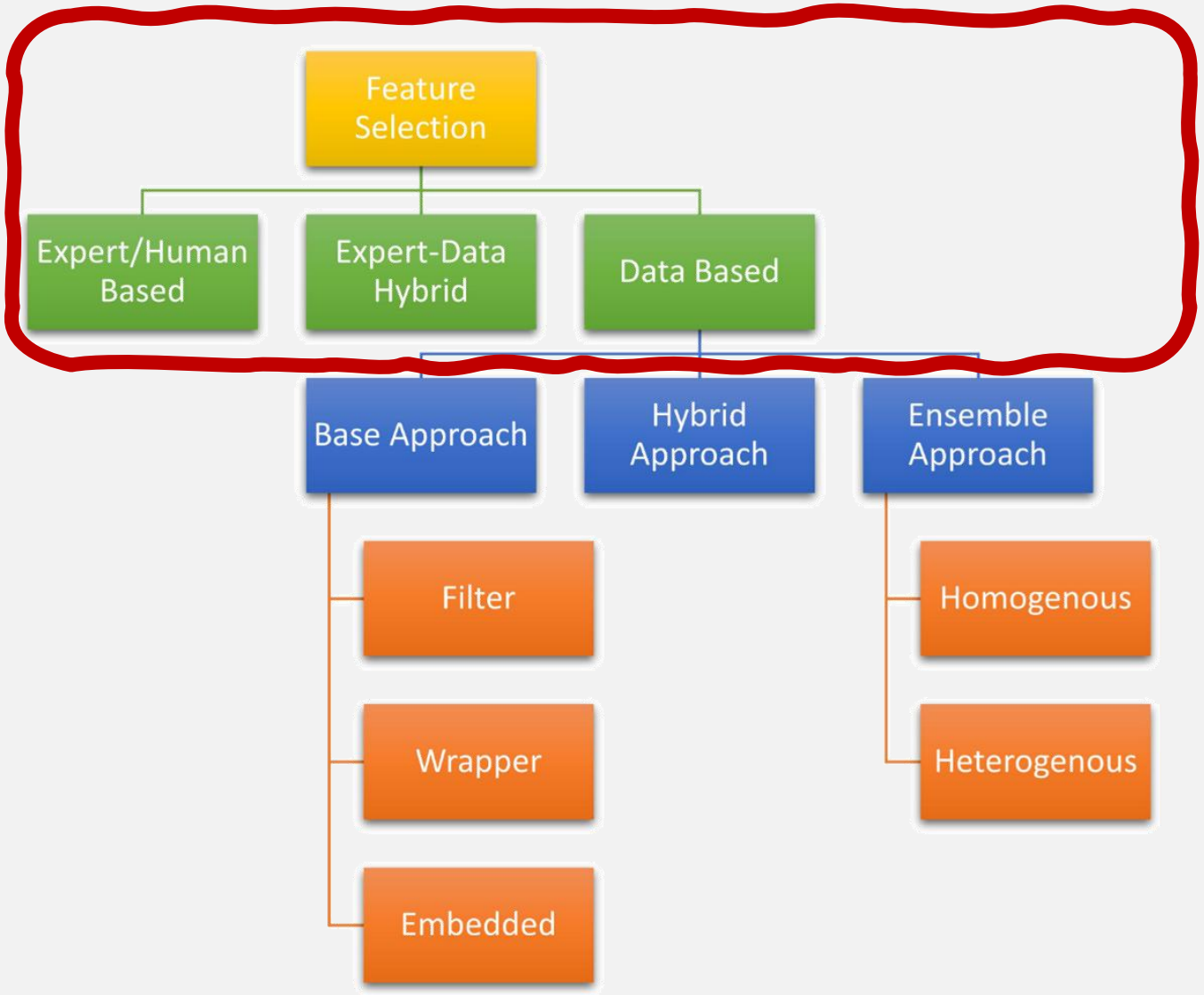
Rahi Jain[1] and Wei Xu[2,*]

[1]*Biostatistics Department, Princess Margaret Cancer Research Centre, Toronto, Ontario, Canada;* [2]*Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada*

1. INTRODUCTION

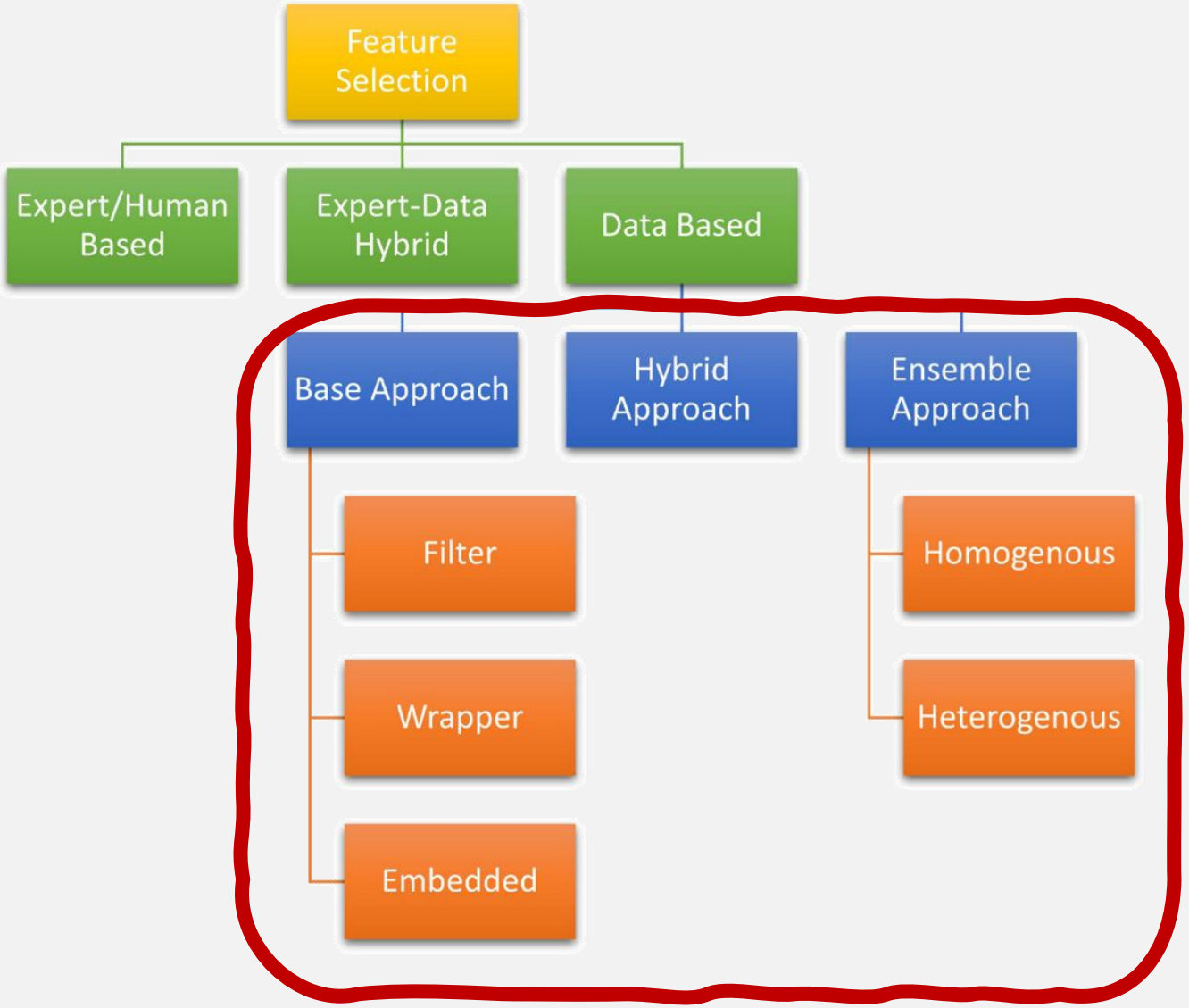
High dimensional data has challenges associated with model fitting, generalizability [1], and computation complexity [2, 3], which prevents modeling by many classic statistical techniques. Feature selection is an important component in high-dimensional data analysis domains like genomics [4] and radiomics [5], as it reduces the dimensions of the dataset. Literature provides many techniques to perform feature selection. However, these techniques could be categorized

Literatura



1. Oparte o wiedzę ekspercką
   - kilka zmiennych
   - brak interakcji
   - dobrze znane w danej dziedzinie
2. Wykorzystuje kombinację wiedzy eksperckiej i informacji z danych
   - uwzględniają wiedzę a priori w procesie selekcji zmiennych
3. Oparte o strukturę danych

Literatura



1. Wykorzystanie pojedynczej metody → pełne dane
   - Wybór w oparciu o strukturę danych np. korelacja
   - Wybór w oparciu o analizę podzbiorów zmiennych
   - Przekształcenie danych
2. Kombinacja wielu metod → pełna dane
3. Selekcja oparta o podzbiory danych

1. Literatura ①
2. **Tradycyjne metody wyboru zmiennych**
3. Regresja z regularyzacją – Penalized regression
4. IC
5. Kryteria dla modeli zagnieżdżonych
6. Literatura ②

Eliminacja wsteczna

model ze wszystkimi zmiennymi

algorytm oparty na P wartościach

1. Test dla istotności każdej zmiennej (np. **t**) → **P wartość**
2. Usunąć zmienną o najwyższej P wartości
3. Iteracja pkt. 1-2
4. Zakończenie – brak nieistotnych zmiennych (np. $P \leq 0.05$)

algorytm oparty na jakości dopasowania modelu

1. Usunąć jedną zmienną
2. Obliczyć jakość dopasowania modelu (np. **R^2**, **AIC**, **deviance**)
3. Trwale usunąć zmienną najmniej pogarszającą jakość dopasowania
4. Iteracja pkt. 1-3
5. Jakość dopasowania poniżej założonego progu (np. $R^2 \leq 50\%$, **P wartość dla deviance** ≥ 0.05)

Eliminacja wsteczna

1. Zalety
   - prosta algorytmicznie implementacja

```
library(stats)
# algorytm oparty na jakości dopasowania modelu = AIC
final_set <- step(lm(y ~ x1 + x2 + x3 + x4),trace = TRUE, direction = "backward")
final_set <- step(glm(y ~ x1 + x2 + x3 + x4, family = binomial) ),trace = TRUE,
direction = "backward")
```

2. Wady
   - konieczność dopasowania pełnego modelu = wszystkie zmienne
   - intensywny obliczeniowo
   - problem wielokrotnego testowania → wiele testów dla zmiennych
   - problemy dla zmiennych skorelowanych
   - Ignoruje interakcje między zmiennymi

Selekcja w przód

model tylko z β_0

algorytm oparty na jakości dopasowania modelu

1. Dodać pojedynczą zmienną
2. Obliczyć jakość dopasowania modelu (np. R^2, AIC, deviance)
3. Pozostawić zmienną o najlepszej jakości dopasowania
4. Iteracja pkt. 1-3
5. Żadna dodana zmienna nie poprawia jakości dopasowania modelu ($\varepsilon \leq 0.01$)

Selekcja w przód

1. Zalety
   - prosta algorytmicznie implementacja

```
library(stats)
# algorytm oparty na jakości dopasowania modelu = AIC
final_set <- step(lm(y ~ x1 + x2 + x3 + x4),trace = TRUE, direction = "forward")
final_set <- step(glm(y ~ x1 + x2 + x3 + x4, family = binomial) ),trace = TRUE,
direction = "forward")
```

2. Wady
   - intensywny obliczeniowo
   - problemy dla zmiennych skorelowanych
   - ignoruje interakcje między zmiennymi
   - **różne, wybrane zestawy zmiennych**

Selekcja krokowa

1. Kombinacja podejścia forward i backward
   - Pierwszy model → pełny / minimalny / „środkowy”
   - Selekcja w przód → eliminacja wsteczna
   - Eliminacja wsteczna → selekcja w przód

Selekcja krokowa

1. Zalety
   - Wykorzystuje zalety obu poprzednich podejść

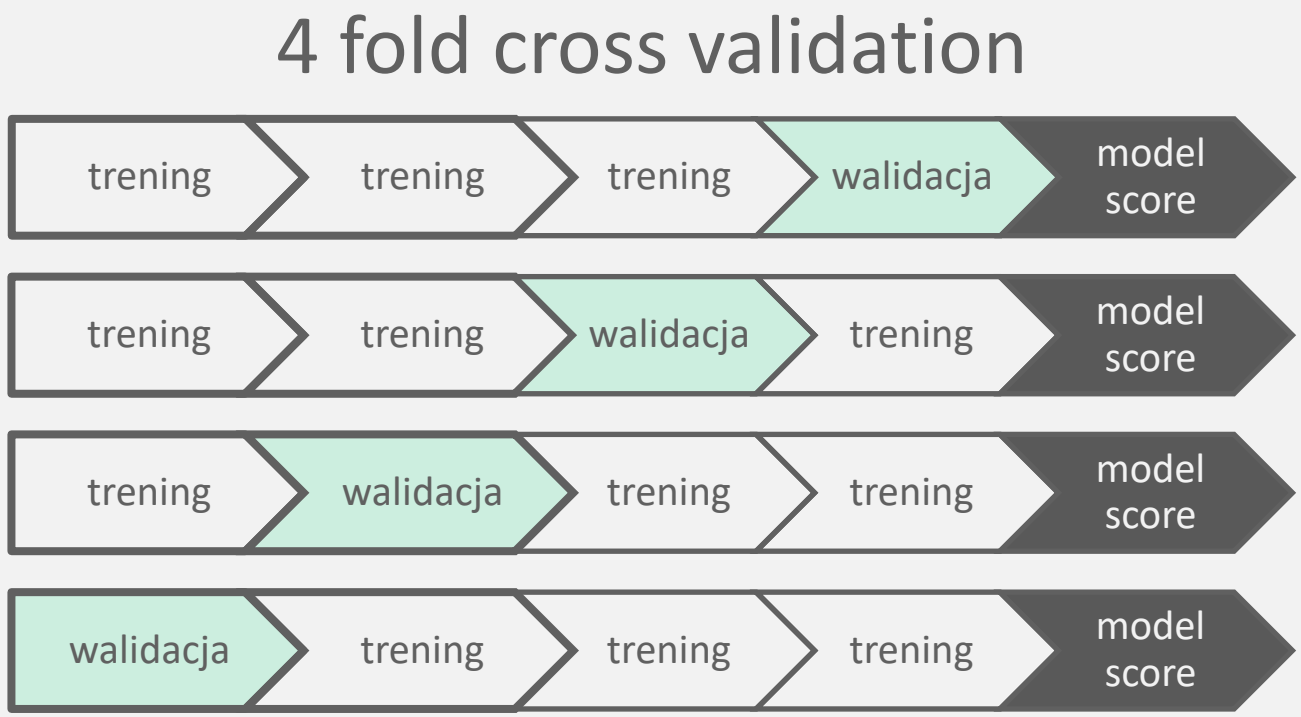
```
library(MASS)
# algorytm oparty na jakości dopasowania modelu = AIC
# można też użyć do selekcji w przód i eliminacji wstecznej → direction
full_model  <- lm(y ~ x1 + x2 + x3 + x4)
final_model <- stepAIC(full_model, direction = "both")
null_model  <- lm(y ~ 1)
final_model <- stepAIC(null_model, direction = "both")

null_model  <- glm(y ~ 1, family = binomial)
final_model <- stepAIC(null_model, direction = "both")
```

2. Wady
   - **Zależny od algorytmu → różne, wybrane zestawy zmiennych**

Walidacja krzyżowa

- Podział dostępnych danych
  - treningowe ①
  - walidacyjne ①
  - treningowe ②
  - walidacyjne ②
  - treningowe ③
  - walidacyjne ③
  - treningowe ④
  - walidacyjne ④
  - testowe
- Walidacja krzyżowa



1. Literatura ①
2. Tradycyjne metody wyboru zmiennych
3. **Regresja z regularyzacją – Penalized regression**
4. IC
5. Kryteria dla modeli zagnieżdżonych
6. Literatura ②

Regularyzacja

- „Kara” za złożoność modelu regresji → liczba efektów = zmiennych niezależnych
  - L1
  - L2
  - AIC
  - BIC

- Pozwala na wybranie optymalnego zestawu zmiennych niezależnych

Regularyzacja

- „Kara” za złożoność modelu regresji → liczba efektów = zmiennych niezależnych
  - L1 $\lambda \sum_{i=1}^{P} \left|\hat{\beta}_i\right|$
  - L2 $\lambda \sum_{i=1}^{P} \hat{\beta}_i^{\,2}$
  - AIC $2P$
  - BIC $\log(N)\, P$

- Pozwala na wybranie optymalnego zestawu zmiennych niezależnych

Regularyzacja Lasso

1. LASSO
   - Regularyzacja L1
   - Least Absolute Shrinkage and Selection Operator
   - Tibshirani (1996)
   - minimum funkcji: $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P}|\hat{\beta}_j|$
   - minimum funkcji: $-\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{p}_i) + (1 - y_i)(1 - \hat{p}_i)] + \lambda \sum_{j=1}^{P}|\hat{\beta}_j|$
2. **Zerowanie estymatorów niektórych zmiennych $\hat{\boldsymbol{\beta}}_j$**
3. Faktyczna selekcja zmiennych → usuwanie zmiennych z modelu przez $\hat{\beta}_j = 0$
4. Problem ze zmiennymi skorelowanymi – dowolnie wybiera jedną zmienną spośród zestawu zmiennych skorelowanych
5. Empiryczny wybór $\lambda \in [0, +\infty]$ np. walidacja krzyżowa, elbow method
6. Estymacja $\hat{\beta}_j$ metodą iteracji

Regularyzacja Lasso

```
library(glmnet)
ibrary(MASS)

# Fit LASSO with CV for best lambda with linear regression
lasso_model <- cv.glmnet(X, y, alpha = 1, standardize=TRUE) # alpha=1 for LASSO
lambda_final <- lasso_model$lambda.min

# Fit LASSO with CV for best lambda with logistic regression
lasso_model <- cv.glmnet(X, y, alpha = 1, standardize=TRUE, family="binomial")
lambda_final <- lasso_model$lambda.min

# Fit LASSO with the final lambda
final_model <- glmnet(X, y, alpha = 1, lambda = lambda_final)
coef(final_model)
```

Regularyzacja Ridge

1. Ridge
   - Regularyzacja L2
   - Hoerl and Kennard (1970)
   - minimum funkcji: $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \boldsymbol{\lambda}\sum_{j=1}^{P}\widehat{\boldsymbol{\beta}}_j^{\,2}$
   - minimum funkcji: $-\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{p}_i) + (1 - y_i)(1 - \hat{p}_i)] \quad + \boldsymbol{\lambda}\sum_{j=1}^{P}\widehat{\boldsymbol{\beta}}_j^{\,2}$
2. **Zmniejszanie wartości estymatorów niektórych zmiennych $\widehat{\boldsymbol{\beta}}_j$**
3. Brak formalnej selekcji zmiennych, wszystkie zmienne pozostają w modelu, również mało istotnie zmienne – zmniejszanie estymatorów $\hat{\beta}_j$
4. Lepiej niż LASSO radzi sobie ze zmiennymi skorelowanymi
5. Lepsza interpretowalność estymatorów
6. Empiryczny wybór $\boldsymbol{\lambda} \in [0, +\infty]$ np. walidacja krzyżowa, elbow method
7. Estymacja $\hat{\beta}_j$ metodą iteracji

Regularyzacja Ridge

```
library(glmnet)
ibrary(MASS)

# Fit Ridge with CV for best lambda with linear regression
ridge_model <- cv.glmnet(X, y, alpha = 0, standardize=TRUE) # alpha=0 for Ridge
lambda_final <- ridge_model$lambda.min

# Fit Ridge with CV for best lambda with logistic regression
ridge_model <- cv.glmnet(X, y, alpha = 0, standardize=TRUE, family="binomial")
lambda_final <- ridge_model$lambda.min


# Fit Ridge with the final lambda
final_model <- glmnet(X, y, alpha = 1, lambda = lambda_final)
coef(final_model)
```

Elastic net

1. Elastic net – kompilacja Lasso i Ridge
   - Regularyzacja L1 i L2
   - minimum funkcji: $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \boldsymbol{\lambda}\left(\boldsymbol{\alpha}\sum_{j=1}^{P}|\hat{\beta}_j| + \sum_{j=1}^{P}\hat{\beta}_j{}^2\right)$
   - minimum funkcji: $-\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{p}_i) + (1 - y_i)(\hat{p}_i)] \quad + \boldsymbol{\lambda}\left(\boldsymbol{\alpha}\sum_{j=1}^{P}|\hat{\beta}_j| + \sum_{j=1}^{P}\hat{\beta}_j{}^2\right)$
   - $\alpha \in [0,1]$ – waga dla regularyzacji L1

2. Łączy zalety podejść LASSO i Ridge – wybór zmiennych i zmniejszanie estymatorów

3. Bardziej wymagające obliczeniowo – estymacja hiperparamterów $\boldsymbol{\lambda}$ i $\boldsymbol{\alpha}$

4. Estymacja $\hat{\beta}_j$ metodą iteracji

Elastic net

```
library(caret)

# Fit Elastic net with CV for best lambda with linear regression
train_control <- trainControl(method = "cv", number = 5)

# Fit Elastic net with CV for best lambda with linear regression
elastic_net_model <- train(response ~ ., data = data,
method = "glmnet",
trControl = train_control,
tuneGrid = expand.grid(alpha = 0:1, lambda = 0:10))

# Fit Elastic net with CV for best lambda with logistic regression
elastic_net_model <- train(response ~ ., data = data,
method = "glmnet",
family = "binomial",
trControl = train_control,
tuneGrid = expand.grid(alpha = 0:1, lambda = 0:10))
```

Adaptive Lasso

1. LASSO z inną wagą dla każdej zmiennej
   - Regularyzacja L1
   - minimum funkcji: $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} w_j \left|\hat{\beta}_j\right|$
   - minimum funkcji: $-\frac{1}{N}\sum_{i=1}^{N}[y_i log(\hat{p}_i) + (1 - y_i)(\hat{p}_i)] \quad + \lambda \sum_{j=1}^{P} w_j \left|\hat{\beta}_j\right|$
   - $w_i = \frac{1}{|\widehat{\beta}_{LS\_j}|^{\gamma}}$
   - $\hat{\beta}_{LS\_j}$ estymator zmiennej zależnej z pełnego modelu regresji (wszystkie zmienne)
   - $\gamma > 0.0$

2. Pełny model musi być estymowalny P < N

3. Estymacja $\hat{\beta}_j$ metodą iteracji

Adaptive Lasso

```
library(glmnet)
ibrary(MASS)

# Fit LASSO with CV for best lambda with linear regression
lasso_model <- cv.glmnet(X, y, alpha = 1, penalty.factor = weights,
standardize=TRUE) # alpha=1 for LASSO
lambda_final <- lasso_model$lambda.min

# Fit LASSO with CV for best lambda with logistic regression
lasso_model <- cv.glmnet(X, y, alpha = 1, penalty.factor = weights,
standardize=TRUE, family="binomial")
lambda_final <- lasso_model$lambda.min

# Fit LASSO with the final lambda
final_model <- glmnet(X, y, alpha = 1, lambda = lambda_final)
coef(final_model)
```

1. Literatura ①
2. Tradycyjne metody wyboru zmiennych
3. Regresja z regularyzacją – Penalized regression
4. **IC**
5. Kryteria dla modeli zagnieżdżonych
6. Literatura ②

AIC

1. Akaike Information Criterion
   - $AIC = N\ log\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N} + 2\ P$ regresja liniowa
   - $AIC = -2\sum_{i=1}^{N}[y_i log(\hat{p}_i) + (1 - y_i)(1 - \hat{p}_i)] + 2\ P$ regresja logistyczna
2. $2\ \mathrm{P}$ – kara za liczbę zmiennych niezależnych w modelu
3. Najlepszy zestaw zmiennych → model z najniższym AIC
4. Metoda nieiteracyjna
5. Możliwe porównanie modeli niezagnieżdżonych
6. Brak istotności statystycznej dla rankingu modeli

BIC

1. Bayesian Information Criterion
   - $BIC = N\ log\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N} + \log(N)\,P$ regresja liniowa
   - $BIC = -2\sum_{i=1}^{N}[y_i log(\hat{p}_i) + (1 - y_i)(1 - \hat{p}_i)] + \log(N)\,P$ regresja logistyczna
2. $\log(N)\,P$ – kara za liczbę zmiennych niezależnych w modelu
3. Najlepszy zestaw zmiennych → model z najniższym BIC
4. Metoda nieiteracyjna
5. Możliwe porównanie modeli niezagnieżdżonych
6. Brak istotności statystycznej dla rankingu modeli
7. BIC nakłada większą karę na liczbę zmiennych niż AIC – $2 < \log(N)$

AICc

1. Corrected Akaike Information Criterion
   - $AICc = N\ log\frac{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}{N} + 2\ P + \frac{2P(P+1)}{N-P-1}$ regresja liniowa
   - $AICc = -2\sum_{i=1}^{N}[y_i log(\hat{p}_i) + (1-y_i)(1-\hat{p}_i)] \quad + 2\ P$ regresja logistyczna
2. kara zależna od N = wielkości próby danych
3. Duża próba $N \to +\infty$, $\frac{2P(P+1)}{N-P-1} \to 0$ wtedy AIC ≈ AICc

mBIC

Extending the Modified Bayesian Information Criterion (mBIC) to Dense Markers and Multiple Interval Mapping FREE

Małgorzata Bogdan ✉, Florian Frommlet, Przemysław Biecek, Riyan Cheng, Jayanta K. Ghosh, R.W. Doerge

1. BIC z różnymi karami dla grup parametrów
2. Stworzony dla kontekstu genetycznego

1. Literatura ①
2. Tradycyjne metody wyboru zmiennych
3. Regresja z regularyzacją – Penalized regression
4. IC
5. **Kryteria dla modeli zagnieżdżonych**
6. Literatura ②

Likelihood ratio test

1. LRT – modele liniowe
   - $LRT = -2\ ln\frac{L_{small}}{L_{large}} = -2ln\frac{\sum_{i=1}^{N}(y_i - \hat{y}_{small\_i})^2}{\sum_{i=1}^{N}(y_i - \hat{y}_{large\_i})^2}$
2. Test !!! → P values dla porównania modeli
3. $LRT \sim \chi^2_{P_{large} - P_{small}}$
4. Hipotezy:
   H_0: duży model nie jest lepszy niż mały model
   H_1: duży model jest lepszy niż mały model
5. Interpretacja → P value = 0.017 ???

Deviance

1. Deviance – modele logistyczne
   - $D = -2\ ln\frac{L_{small}}{L_{large}} = -2\sum_{i=1}^{N}\left[y_i log\left(\frac{\hat{p}_{large\_i}}{\hat{p}_{small\_i}}\right) + (1-y_i)\left(\frac{1-\hat{p}_{large\_i}}{1-\hat{p}_{small\_i}}\right)\right]$
2. Test → P values dla porównania modeli
3. $D \sim \chi^2_{P_{large}-P_{small}}$
4. Hipotezy:
   H_0: duży model nie jest lepszy niż mały model
   H_1: duży model jest lepszy niż mały model
5. Interpretacja → P value = 0.276 ???

1. Literatura ①
2. Tradycyjne metody wyboru zmiennych
3. Regresja z regularyzacją – Penalized regression
4. IC
5. Kryteria dla modeli zagnieżdżonych
6. **Literatura ②**

Literatura

Feature Selection Strategies for Deep Learning-Based Classification in Ultra-High-Dimensional Genomic Data

by **Krzysztof Kotlarz** [1], **Dawid Słomian** [2], **Weronika Zawadzka** [1] and **Joanna Szyda** [1,2,*]

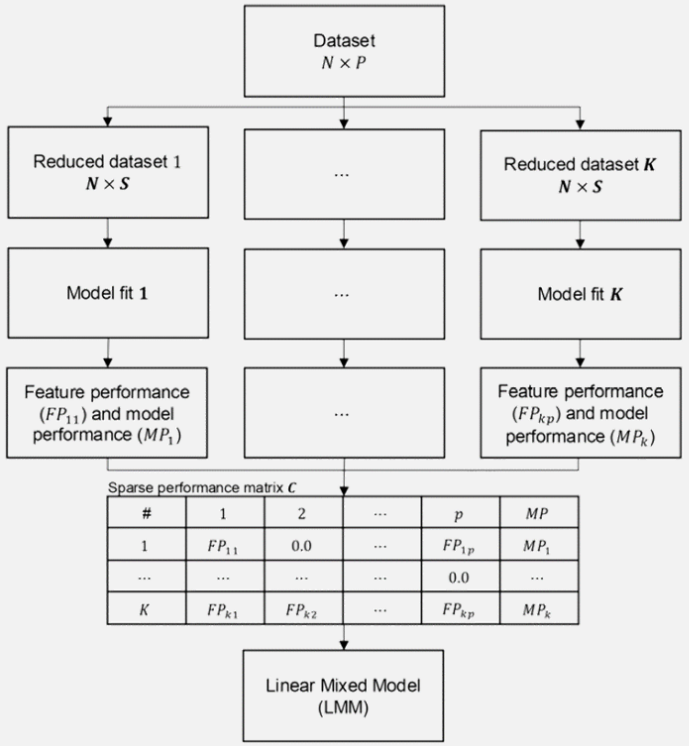
[1] Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, 51-631 Wroclaw, Poland

[2] National Research Institute of Animal Production, 32-083 Balice, Poland

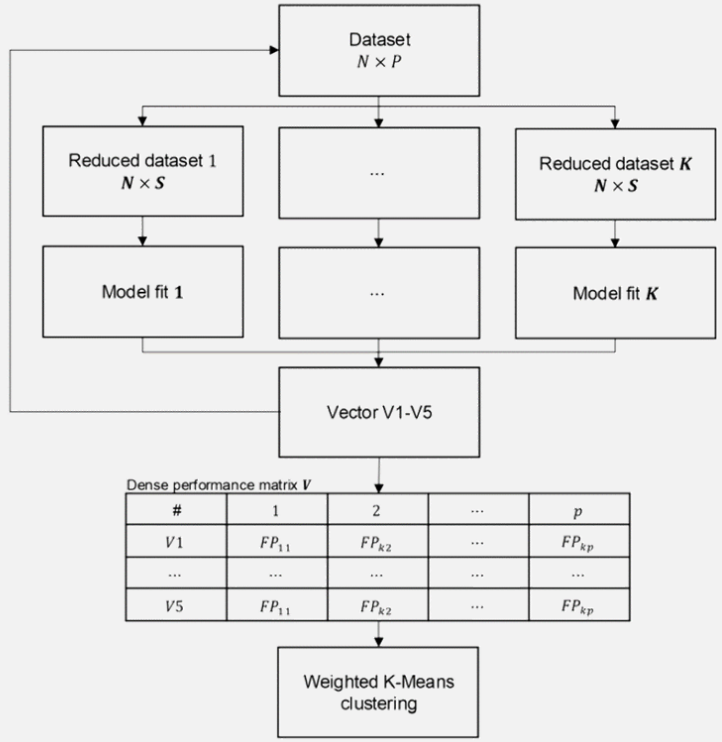
[*] Author to whom correspondence should be addressed.

Int. J. Mol. Sci. **2025**, *26*(16), 7961; https://doi.org/10.3390/ijms26167961

A



B



1. Literatura ①
2. Tradycyjne metody wyboru zmiennych
3. Regresja z regularyzacją – Penalized regression
4. IC
5. Kryteria dla modeli zagnieżdżonych
6. Literatura ②

1. Po co jest potrzebny wybór zmiennych w modelu ?
2. Które z metod regularyzacyjnych jest najlepsze?
3. Co to jest IC?
4. Czy przy użyciu AIC można porównać modele zagnieżdżone?