DREAM5 Challenges

Metody i rezultaty

Julia Herman-Iżycka Jacek Jendrej

Praktyki wakacyjne 2010 – sesja sprawozdawcza

Plan prezentacji

Plan prezentacji

Plan prezentacji

Plan prezentacji

Czym jest *Machine learning*

Terminologia

- $\left(x^{(i)}, y^{(i)}\right) \in \mathcal{X} \times \mathcal{Y}, i \in \{1, \ldots, m\}$ – egzemplarz treningowy
- $y^{(i)} \in \mathcal{Y}$ – zmienna wyjściowa, cel
- Poszczególne współrzędne wektorów $x^{(i)}$ to tzw. atrybuty.
- Wynikiem algorytmu jest hipoteza h, czyli po prostu funkcja $h : \mathcal{X} \to \mathcal{Y}$. Wybieranie h nazywamy trenowaniem.
- Algorytm testujemy, wybierając (niewielki) podzbiór $T \subset \{1, \ldots, m\}$, pomijając egzemplarze o indeksach z T podczas trenowania, a następnie porównując $h\left(x^{(i)}\right)$ z $y^{(i)}$ dla $i \in T$.

Challenge 1 - Epitope-Antibody Recognition

- Przeciwciała i ich paratopy
- Epitopy - **sekwencyjne** i przestrzenne
- Szukamy sekwencyjnych epitopów wiążących się do przeciwciał IVIg

Challenge 1 - Epitope-Antibody Recognition

Dane

Tabela 13,638 sekwencji aminokwasowych długości 13 - 21 + ich siły wiązania wyrażone liczbami z zakresu 1-65,423.

- siła wiązania $>$ 10.000 to sekwencje pozytywne - 3,420 peptydów
- siła wiązania $<$ 1.000 to sekwencje negatywne - pozostałe 10,218

Problem

Klasyfikacja 13,640 sekwencji długości 9 - 20 aminokwasów. Należy podać pewność zaklasyfikowania sekwencji jako pozytywna.

Challenge 2 - TF-DNA Motif Recognition

Czynniki transkrypcyjne

Czynniki transkrypcyjne oddziaływują z DNA regulując w ten sposób transkrypcję. Zwykle rozpoznają motywy długości 6-12 nukleotydów.

Co chcemy zrobić?

Chcemy przewidywać siłę wiązania danego czynnika transkrypcyjnego do danej sekwencji, na podstawie wiązania się tego czynnika do reprezentatywnego zbioru sekwencji

Challenge 2 - TF-DNA Motif Recognition

Sekwencje

Użyte zostały dwa zbiory sekwencji, każdy liczący ok 40.000 sekwencji długości 35. Zbiory zostały stworzone na podstawie ciągów de Bruijn'a, każdy zawiera wszystkie możliwe 10-mery oraz po 32 egzemplarze każdego niepalindromicznego 8-meru.

Dane

Dla każdego czynnika transkrypcyjnego otrzymaliśmy średnie siły sygnału dla sekwencji z jednej z grup.

Problem

Należy przewidzieć siłę wiązania 66 czynników transkrypcyjnych do sekwencji z jednej grupy, na podstawie wiązania sekwencji z drugiej grupy.

Adaptujemy algorytm Smith'a-Waterman'a

Algorytm Smith'a-Waterman'a

Algorytm Smith'a-Waterman'a jest dynamicznym algorytmem znajdującym najlepsze lokalne uliniowienie dwóch sekwencji. Korzystaliśmy z implementacji na karty graficzne. Jako macierzy podobieństwa użyliśmy kolejno macierzy BLOSUM80, BLOSUM62 oraz BLOSUM45.

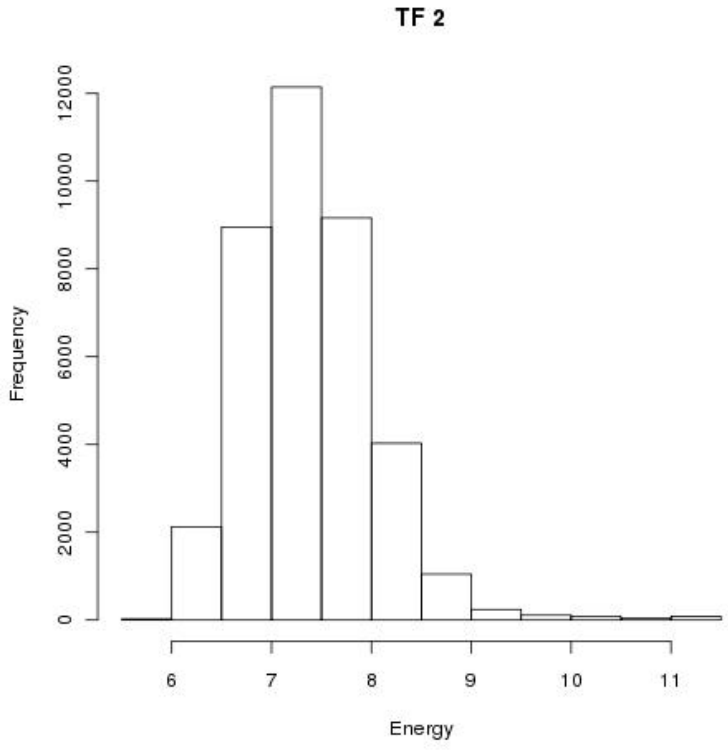
Wykorzystanie

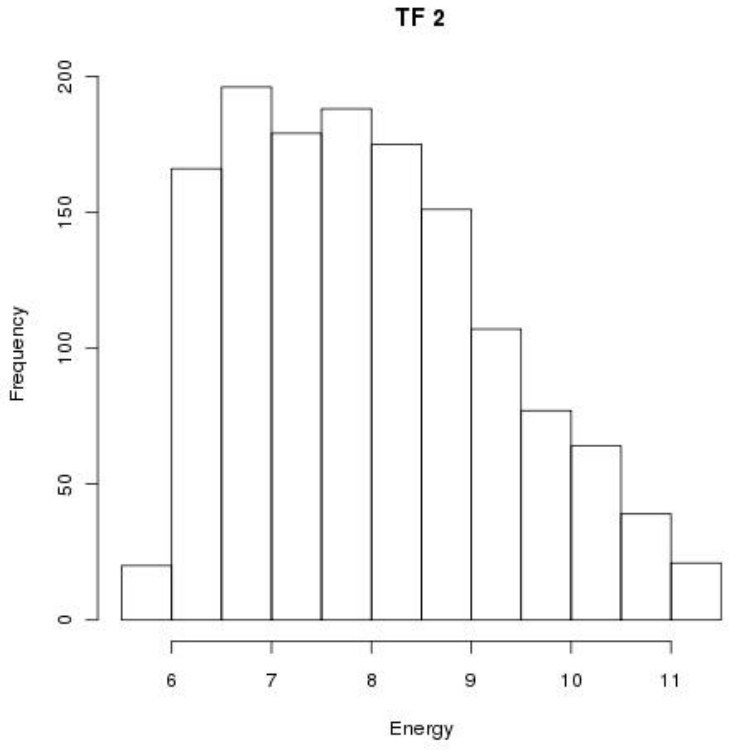
Jako atrybutów użyliśmy maksymalnych podobieńst sekwencji ze zbioru testowego do pozytywnych sekwencji ze zbioru treningowego.

Sami poszukajmy motywów

- Dla przeciwciał struktura drugorzędowa jest kluczowa.
- Ustalamy indeks i szukamy trzyresiduowych motywów – kolejnych aminokwasów, albo co drugi (β) lub co trzeci (α).
- Podobieństwo sekwencji do motywu określa minimalna odległość euklidesowa od fragmentu sekwencji (względem danego indeksu).
- Czasem się udaje – wśród 100 najbliższych sekwencji zdarzało się ponad 70 wiążących, mimo że ogółem jest ich trzykrotnie mniej niż niewiążących.
- Wybieramy skuteczne motywy; odległości od nich tworzą atrybuty.

Zbalansujmy dane





Pomijamy całkowicie większość danych! Ale za to mamy równomierną reprezentację wszystkich przedziałów energetycznych.

n-gramy

- Każdy z 4^n n-zasadowych łańcuchów tworzy atrybut – T jeśli dany łańcuch występuje jako podsłowo w sekwencji, F w przeciwnym wypadku.
- Trenujemy las.
- Używamy $n = 4, 5$. $n = 6$ – przeuczenie, za mało danych.
- Boruta wybiera ważne atrybuty. To prawdopodobnie nie poprawia skuteczności.

n-gramy

- Każdy z 4^n n-zasadowych łańcuchów tworzy atrybut – T jeśli dany łańcuch występuje jako podsłowo w sekwencji, F w przeciwnym wypadku.
- Trenujemy las.
- Używamy $n = 4, 5$. $n = 6$ – przeuczenie, za mało danych.
- Boruta wybiera ważne atrybuty. To prawdopodobnie nie poprawia skuteczności.

n-gramy

- Każdy z 4^n n-zasadowych łańcuchów tworzy atrybut – T jeśli dany łańcuch występuje jako podsłowo w sekwencji, F w przeciwnym wypadku.
- Trenujemy las.
- Używamy $n = 4, 5$. $n = 6$ – przeuczenie, za mało danych.
- Boruta wybiera ważne atrybuty. To prawdopodobnie nie poprawia skuteczności.

n-gramy

- Każdy z 4^n n-zasadowych łańcuchów tworzy atrybut – T jeśli dany łańcuch występuje jako podsłowo w sekwencji, F w przeciwnym wypadku.
- Trenujemy las.
- Używamy $n = 4, 5$. $n = 6$ – przeuczenie, za mało danych.
- Boruta wybiera ważne atrybuty. To prawdopodobnie nie poprawia skuteczności.

n-gramy

- Każdy z 4^n n-zasadowych łańcuchów tworzy atrybut – T jeśli dany łańcuch występuje jako podsłowo w sekwencji, F w przeciwnym wypadku.
- Trenujemy las.
- Używamy $n = 4, 5$. $n = 6$ – przeuczenie, za mało danych.
- Boruta wybiera ważne atrybuty. To prawdopodobnie nie poprawia skuteczności.

Rank++

Dla każdego czynnika transkrypcyjnego użyliśmy programu RankMotif++ do znalezienia motywu długości 8 w sekwencjach treningowych.
Następnie przy pomocy programu MAST z pakietu MEME dopasowaliśmy sekwencje do znalezionych motywów. Użyliśmy p-value tego dopasowania jako atrybut.

Skuteczność klasyfikatorów składowych w EAR

Współczynnik korelacji (Pearsona)

Klasyfikator	z en.	z dec.	DecAcc
bayes	0.4995	0.5284	1.0627
kNN	0.3820	0.4008	0.8705
witold.forest	0.4754	0.4959	0.5570
witold.forest	0.5208	0.5461	1.1039
ngram.forest	0.4330	0.4606	0.3315
similarity	0.4004	0.4068	1.3115

Klasyfikator	TPV	TNV
ngram.forest	0.7480	0.7636
witold.forest	0.7784	0.7726

Skuteczność klasyfikatorów składowych w EAR

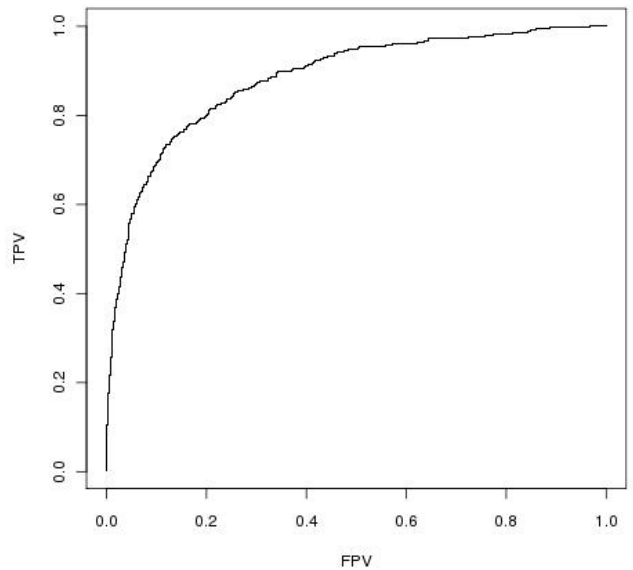
Współczynnik korelacji (Pearsona)

Klasyfikator	z en.	z dec.	DecAcc
bayes	0.4995	0.5284	1.0627
kNN	0.3820	0.4008	0.8705
witold.forest	0.4754	0.4959	0.5570
witold.forest	0.5208	0.5461	1.1039
ngram.forest	0.4330	0.4606	0.3315
similarity	0.4004	0.4068	1.3115

Klasyfikator	TPV	TNV
ngram.forest	0.7480	0.7636
witold.forest	0.7784	0.7726

Skuteczność końcowa w EAR

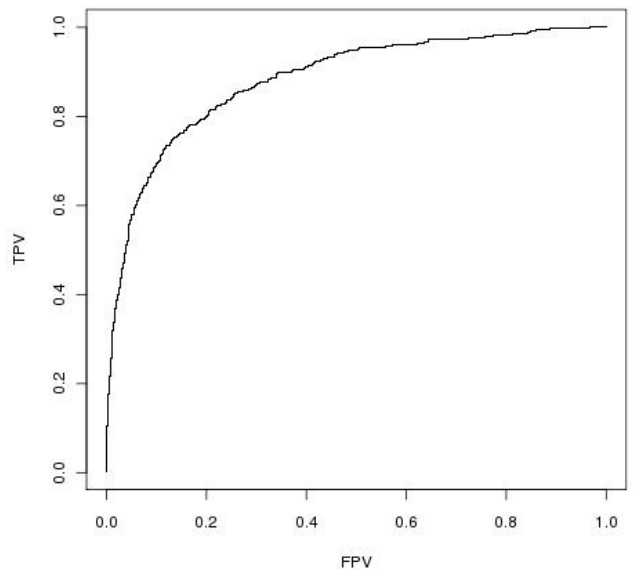
- TPV = 0.792, TNV = 0.812
- Proporcja głosów ∼ pewność decyzji
- ROC:



- AUC ≃ 0.883

Skuteczność końcowa w EAR

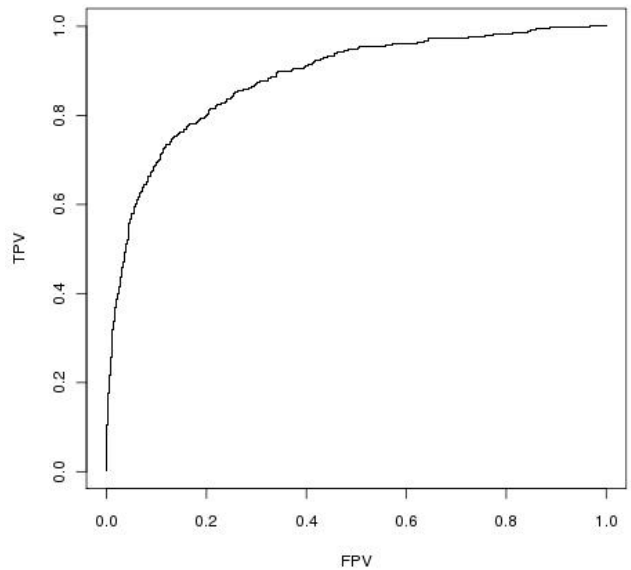
- TPV = 0.792, TNV = 0.812
- Proporcja głosów ∼ pewność decyzji
- ROC:



- AUC ≃ 0.883

Skuteczność końcowa w EAR

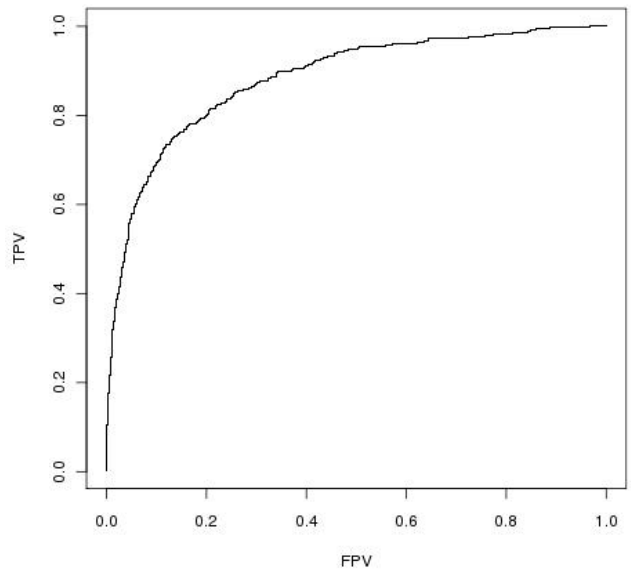
- TPV = 0.792, TNV = 0.812
- Proporcja głosów ∼ pewność decyzji
- ROC:



- AUC ≃ 0.883

Skuteczność końcowa w EAR

- TPV = 0.792, TNV = 0.812
- Proporcja głosów $\sim$ pewność decyzji
- ROC:



- AUC $\simeq$ 0.883

Skuteczność klasyfikatorów składowych w PBM

Współczynnik korelacji (Pearsona)

TF	4-gram	Bor	Atr	5-gram	Bor	Atr	mast
1	0.63	0.59	48	0.64	0.53	54	-0.43
25	0.80	0.79	32	0.85	0.86	53	-0.66
41	0.63	0.60	42	0.64	0.59	58	-0.44
Śr.	0.73	0.70	45.2	0.75	0.70	62.3	-0.49

Procent wyjaśnionej wariancji (RSQ)

TF	4-gram	Bor	Atr	5-gram	Bor	Atr
1	0.37	0.34	48	0.38	0.25	54
25	0.62	0.61	32	0.72	0.74	53
41	0.34	0.35	42	0.40	0.33	58
Śr.	0.51	0.51	45.2	0.55	0.50	62.3

Skuteczność klasyfikatorów składowych w PBM

Współczynnik korelacji (Pearsona)

TF	4-gram	Bor	Atr	5-gram	Bor	Atr	mast
1	0.63	0.59	48	0.64	0.53	54	-0.43
25	0.80	0.79	32	0.85	0.86	53	-0.66
41	0.63	0.60	42	0.64	0.59	58	-0.44
Śr.	0.73	0.70	45.2	0.75	0.70	62.3	-0.49

Procent wyjaśnionej wariancji (RSQ)

TF	4-gram	Bor	Atr	5-gram	Bor	Atr
1	0.37	0.34	48	0.38	0.25	54
25	0.62	0.61	32	0.72	0.74	53
41	0.34	0.35	42	0.40	0.33	58
Śr.	0.51	0.51	45.2	0.55	0.50	62.3

Skuteczność końcowa w PBM

- Wielkość zbioru treningowego od 403 do 842, średnio 621.5
- RSQ:

TF	Rozmiar	4-gram	5-gram	Blending
1	561	0.37	0.38	0.41
25	707	0.62	0.72	0.75
41	769	0.34	0.40	0.41
Śr.	621.5	0.51	0.55	0.59

- Korelacja RSQ z rozmiarem zbioru treningowego (aż?) 0.34

Skuteczność końcowa w PBM

- Wielkość zbioru treningowego od 403 do 842, średnio 621.5
- RSQ:

TF	Rozmiar	4-gram	5-gram	Blending
1	561	0.37	0.38	0.41
25	707	0.62	0.72	0.75
41	769	0.34	0.40	0.41
Śr.	621.5	0.51	0.55	0.59

- Korelacja RSQ z rozmiarem zbioru treningowego (aż?) 0.34

Skuteczność końcowa w PBM

- Wielkość zbioru treningowego od 403 do 842, średnio 621.5
- RSQ:

TF	Rozmiar	4-gram	5-gram	Blending
1	561	0.37	0.38	0.41
25	707	0.62	0.72	0.75
41	769	0.34	0.40	0.41
Śr.	621.5	0.51	0.55	0.59

- Korelacja RSQ z rozmiarem zbioru treningowego (aż?) 0.34