Podstawy teorii uczenia maszynowego

Wyklad 1

Mateusz Serocki

Politechnika Gdańska

March 16, 2024

Agenda

Kilka słow o mnie

- Aktualnie starszy inżynier uczenia maszynowego w NIKE
- Main skillset: Python/AWS
- 6 lat doświadczenia zawodowego
- Main topics: Regresja, Klasyfikacja, Prognoza, MLOps
- LinkedIn: https://www.linkedin.com/in/mateuszserocki/
- Email: Mateusz.Serockiii@gmail.com

Definicja

Uczenie maszynowe, samouczenie sie maszyn albo systemy uczace sie (ang. machine learning) – obszar sztucznej inteligencji poświecony algorytmom, które poprawiaja sie automatycznie poprzez doświadczenie[1], czyli ekspozycje na dane. Algorytmy uczenia maszynowego buduja model matematyczny na podstawie przykładowych danych, zwanych zbiorem uczacym, w celu prognozowania lub podejmowania decyzji bez bycia zaprogramowanym explicite przez człowieka do tego celu. Algorytmy uczenia maszynowego sa wykorzystywane w wielu różnych zastosowaniach, takich jak ochrona przed spamem (filtrowanie wiadomości internetowych pod katem niechcianej korespondencji), czy rozpoznawanie obrazów, w których opracowanie konwencjonalnych algorytmów do wykonywania potrzebnych zadań jest trudne lub niewykonalne.

Supervised Learning

Typ problemu w którym model uczy sie na danych wejściowych X1 przewidywać target Y1, nastepnie wykorzystuje nowe dane X2, do predykcji wartosci Y2. Tego typu problem dzieli sie na dwie kategorie, w zależności czym jest Y.

Regression

Regresja jest wykorzystywana wtedy kiedy nasza zmienna Y jest ciagła, przykładem może być cena mieszkania.

Classification

Klasyfikacja jest wykorzystywana jeżeli nasza zmienna jest binomialna (ma różne klasy wartości), przykładem może być problem binarny (0,1) np kupił/nie kupił, zdał/nie zdał, ale może to być również problem wieloklasowy np ocena z kolokwium 2/3/4/5 itd.

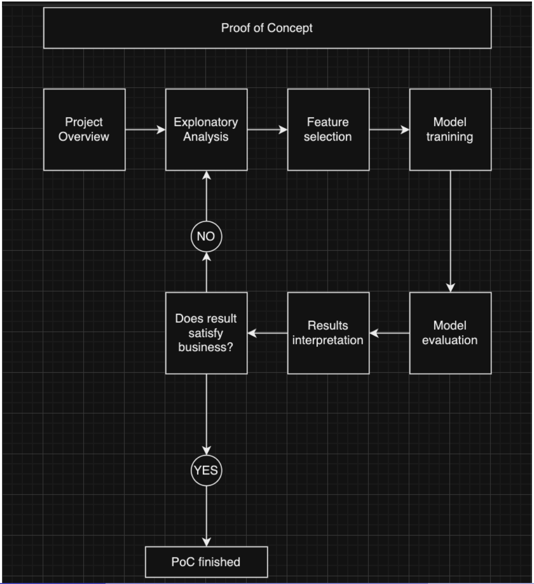
Unsupervised Learning

Typ problemu w którym model uczy sie na danych wejściowych X i na tej podstawie próbuje dostrzec relacje miedzy obserwacjami. Kończy sie to zgrupowaniem obserwacji w tzw. cluster, tego typu algorytmy nazywamy potocznie algorytmami klastrujacymi lub grupujacymi. Przykładem tego typu problemu może być analiza modeli podobnych wśród ubrań, lub grupowanie uczniów na podstawie ich ocen i frekwencji

Reinforcement Learning

Typ problemu w którym model sie pewnego problemu na podstawie wykonywania operacji, wraz z wykonywanymi operacjiami model rozumie czy wykonuje te operacje poprawnie badź nie. Przykład algorym do gry w szachy, wychodzenie z labiryntu itd.

Przykład architektury rozwiazania dla problemu uczenia maszynowego



Pobranie danych

Każdy algorym omawiany na tych zajeciach bedzie wymagał danych, skad takie dane brać oraz w jakim formacie takie dane moga być przechowywane?

Źródła danych

Dane moga pochodzić z różnych źródeł, jednego lub kilku jednocześnie, przykłady źródeł danych

- dane z pliku (.csv, .pdf, .xlsx itd.)
- z bazy danych (postgress, mongodb itd.)
- z internetu (web scrapping)
- chmura (AWS/GCP/Azure)

Przygotowanie danych

Najcześciej dane które otrzymamy należy przetworzyć, w skład przetwarzania danych wchodzi

- czyszczenie danych
- tworzenie nowych zmiennych (np One-Hot encoding)
- standaryzacja danych
- analiza wartości odstajacych

Czyszczenie danych

Jednym z pierwszych kroków które powinniśmy podjać na poczatku pracy z danymi jest ich czyszczenie, warto zwrócic uwage czy nasze dane nie zawieraja błedów. Idealnie czy wszyskie wartości w jednej zmiennej posiadaja ten sam TYP, np. czy jeżeli cecha to wiek i wiekszość obserwacji to liczby, to czy aby na pewno nie ma tam stringów (słow/znaków).

Tworzenie nowych zmiennych

Warto sie zastanowić czy nasze zmienne możemy wyrazić w inny sposób, np poprzez zsumowanie kilku cech, zastapienia kilku cech jedna cecha, podniesienie do kwadratu jeden z cech. Dodanie flagi do poszczególnych obserwacji.

Reduckja wymiarowości

Definicja

Redukcja wymiarowości to transformacja danych z przestrzeni wielowymiarowej do przestrzeni niskowymiarowej, tak aby reprezentacja niskowymiarowa zachowała pewne znaczace właściwości oryginalnych danych, idealnie zbliżone do ich wewnetrznego wymiaru.

Interpretacja słowna

Innymi slowym staramy sie zmniejszyć wymiarowość naszego zbioru danych przy równoczesnym zachowaniu maksymalnej ilości informacji jaka z tego zbioru pochodzi.

Przykład redukcji wymiarow

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0.0
1	4.9	3.0	1.4	0.2	0.0
2	4.7	3.2	1.3	0.2	0.0
3	4.6	3.1	1.5	0.2	0.0
4	5.0	3.6	1.4	0.2	0.0
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	2.0
146	6.3	2.5	5.0	1.9	2.0
147	6.5	3.0	5.2	2.0	2.0
148	6.2	3.4	5.4	2.3	2.0
149	5.9	3.0	5.1	1.8	2.0

Figure: Full dataset

	feature1	feature2	target
0	5.80	5.20	0.0
1	5.60	5.00	0.0
2	5.35	4.80	0.0
3	5.35	4.70	0.0
4	5.70	5.10	0.0
...	...	...	...
145	9.30	7.85	2.0
146	8.80	7.25	2.0
147	9.10	7.50	2.0
148	8.90	7.35	2.0
149	8.45	6.80	2.0

150 rows × 3 columns

Figure: Reduced dataset

Wady redukcji wymiarów

Redukcja wymiarów ma też negatywne skutki takie jak:

- Zmniejszenie ilości informacji
- Ryzyko usuniecia potencjalnie informatywnej zmiennej
- Calkowita lub cześciowa utrata interpretowalności zmiennych
- Kolejny krok potrzebny do przygotowania danych

Zalety redukcji wymiarów

Pozytywne aspekty redukcji wymiarów to:

- Redukcja szumu pochodzacego ze zmiennych
- Brak zmiennych które nie maja uzasadnionego wpływu na wynik
- Mniejsza moc obliczeniowa potrzebna do utworzenia modelu
- Mozemy zrezygnowac z pobierania niektorych zmiennych jezeli na tym etapie sadzimy ze sa nieistotne

Metody reduckji wymiarowosci

Trzy podstawowe metody redukcji wymiarowości, które należy rozważyć to

- Analiza wariancji
- Analiza korelacji
- Braki danych
- PCA
- LASSO (*zostanie omówione przy modelach liniowych z regularyzacja)
- Analiza informatywności cech (*zostanie omówione w rozdziale dot. modeli XGBoost)

Wariancja

Teoria

W teorii prawdopodobieństwa i statystyce wariancja jest kwadratem odchylenia od średniej zmiennej losowej. Wariancje czesto definiuje sie także jako kwadrat odchylenia standardowego.

Zastosowanie

W praktyce wariancja odzwierciedla zmienność danej cechy, niska wariancja może nieść za soba niska informatywność. Warto roważyć usuniecie takiej cechy jeżeli nasz zbiór jest bardzo duży. Koniecznie musimy sprawdzić wpływ tego usuniecia na wynik modelu.

Korelacja

Typy korelacji

- Pearson - standardowa korelacja liniowa
- Spearman - korelacja rankingowa
- Kendal - korelacja rankingowa dla grup z powtarzajacymi sie wartosciami
- more...

Przykład liczenia korelacji Pearsona

Link

Teoria dot. korelacji spearmana

Link

Teoria dot. korelacji Kendalla

Link

Braki danych

Sposoby na radzenie sobie z brakiem danych

- Usuniecie
- Inputacja
- Podstawienie

Przyklad 1

Przykładem algorytmu który radzi sobie z brakami danych jest XGBoost który wykorzystuje średnia wartość danej zmiennej zamiast wartości NULL (brak)

Przyklad 2

Przykładem algorytmu który nie radzi sobie z brakami danych jest regresja liniowa (i pochodne) które w przypadku napotkania wartości null zwróca bład processowania

Standaryzacja zmiennych

Zapamietaj

Każdy algorytm który działa na podstawie liczenia odległości miedzy różnymi zmiennymi wymaga od nas wykonanai standaryzacji

Typy standaryzacji

- MinMaxScaler
- StandardScaler

Przykład

Na wykładzie.

Pytanie z *

Czy w przypadku gdy wszystkie nasze cechy posiadaja rozkład binarny, standaryzacja jest wymagana?

Principal Component Analysis

PCA

Analiza głównych składowych (PCA) to popularna technika analizy dużych zbiorów danych zawierajacych duża liczbe wymiarów/cech na obserwacje, zwiekszajaca interpretowalność danych przy jednoczesnym zachowaniu maksymalnej ilości informacji i umożliwiajaca wizualizacje danych wielowymiarowych

Przyklad numeryczny

Na wykładzie przejdziemy przez przykład numeryczny, podsumowanie dostepne pod Link

Przykład redukcji wymiarow

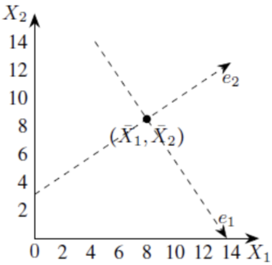


Figure: Rzut wektorów wlasnych

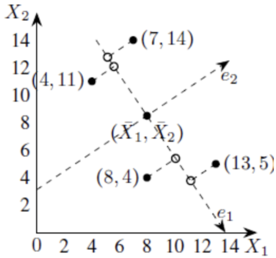


Figure: Mapowanie na podstawie wektorów wlasnych

Wykrywanie wartości odstajacych

Warto zwrócic uwage, że model uczy sie na danych wyjściowych i stara sie optymalizować problem globalnie, jeżeli w naszych danych beda duże wartości odstajace może to znaczaco wpłynać na wyniki naszego modelu. Do analizy czy nasza cecha ma wartości odstajace możemy użyć np box-plota.

Box plot

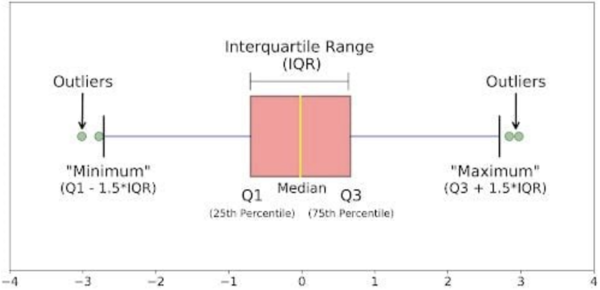


Figure: Box plot - example

Trenowanie modelu

W ML istnieje cała gama modeli, które możemy wytrenować. Dobieramy model zależnie do problemu, innego modelu użyjemy do klassyfikacji, innego do regressji natomiast innego do forecastu. Na poczatku należy określić problem z jakim mamy doczynienia, czy naszym rezultatem ma byc predykcja, a może grupowanie zmiennych?

RandomForest/Drzewa decyzyjne

Opis algorytmu

Losowe lasy lub losowe lasy decyzyjne to metoda uczenia sie zespołowego do klasyfikacji, regresji i innych zadań, która polega na konstruowaniu wielu drzew decyzyjnych w czasie szkolenia.

Zastosowanie

- Regresja
- Klasyfikacja

Dokładne dzialanie algorytmu

Algorytm omówiony na wykładzie, podsumowanie dostepne pod Link

RandomForest

Feature Importance

- Gini Importance / Mean Decrease in Impurity (MDI)
- Permutation Importance or Mean Decrease in Accuracy (MDA)

Omówienie metod

Metody omówione na wykładzie, podsumowanie dostepne pod Link

XGBoost

Opis

XGBoost jest zbiorowym, bazujacym na drzewach, algorytmem uczenia maszynowego, wykorzystujacym strukture wzmacniajaca gradient

Przyklad dzialania algorytmu

Omwówiony na wykładzie, Link

DBSCAN

DBSCAN (od ang. Density-Based Spatial Clustering of Applications with Noise) – algorytm grupowania danych (klasteryzacji) oparty na gestości. W algorytmie używamy dwóch parametrów.

- epsilon
- minpts

DBSCAN

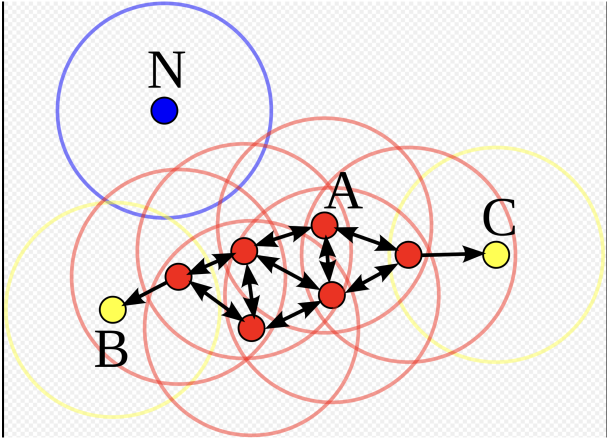


Figure: DBSCAN

Model liniowy

Prosta regresja liniowa (univariate or simple linear regression) polega na przewidywaniu odpowiedzi Y (zmiennej zależnej) na podstawie pojedynczej zmiennej niezależnej X. Regresja liniowa zakłada, że pomiedzy X i Y istnieje zależność liniowa, tzn taka która da sie opisać funkcja liniowa:

$$Y = \beta_0 + \beta_1 * X \tag{1}$$

β_0, β_1 nazywane sa parametrami modelu. Trenowanie modelu polega na estymacji tych parametrów, a nastepnie na postawie wartości x przewidzenie ŷ

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1} * X_1 \tag{2}$$

Metoda najmniejszych kwadatów

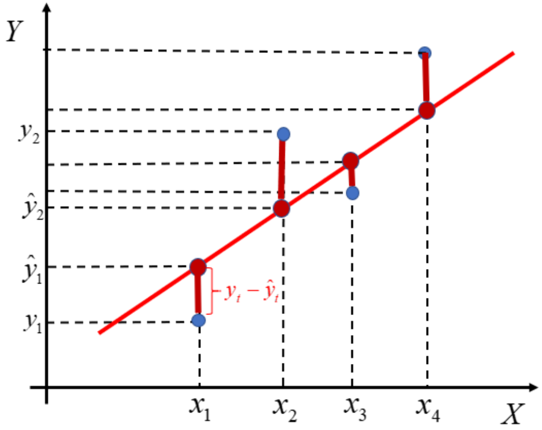


Figure: Metoda najmniejszych kwadratów

Funkcja straty

Jeżeli

$$Y = \beta_0 + \beta_1 * X \tag{3}$$

jest naszym modelem to mozemy okreslic funkcje straty jako

$$MSE = \frac{\sum_{k=1}^{N} (y - \hat{y})^2}{N} \tag{4}$$

a N to liczba obserwacji. Naszym zadaniem jest znalezienie jak najmniejszej funkcji straty. W przypadku mean squaraed error rozwiazanie jest deterministyczne.

Rodzina wykładnicza

Definition

Rodzina rozkładów prawdopodobieństwa nazywa sie rodzina wykładnicza, jeżeli każdy należacy do niej rozkład ma funkcje rozkładu prawdopodobieństwa postaci

$$f(y|\theta, \phi) = exp\left\{\frac{y \cdot \theta - b(\theta)}{\phi} + c(y, \phi)\right\},$$

gdzie:
θ – parametr kanoniczny ($\theta \in \mathbf{R}$),
ϕ – parametr dyspersji ($\phi \in \mathbf{R}_+$),
$b(\theta)$ – jest funkcja dwukrotnie różniczkowalna z dodatnia druga pochodna,
$c(y, \phi)$ – jest funkcja niezależna od parametru θ oraz $y \in \mathbf{R}$.

Definition

Funkcja prawdopodobieństwa rozkładu binarnego przedstawia sie nastepujaco:

$$f(y) = \begin{cases} \mu & \text{gdy } y = 1 \\ 1 - \mu & \text{gdy } y = 0. \end{cases}$$

Wartość oczekiwana oraz wariancja dla zmiennej losowej Y o rozkładzie binarnym wynosi:

$$E(Y) = \mu$$

$$Var(Y) = \mu(1 - \mu).$$

Rozkład binarny należy do rodziny wykładniczej.

Proof.

Funkcje prawdopodobieństwa rozkładu binarnego możemy zapisać za pomoca rodziny wykładniczej:

$$
\begin{gathered}
f(y) = \mu^y \cdot (1-\mu)^{1-y} = \\
= exp\left\{ y \cdot log\,\mu + (1-y) \cdot log(1-\mu) \right\} = \\
= exp\left\{ y \cdot log\,\mu + log(1-\mu) - y \cdot log(1-\mu) \right\} = \\
= exp\left\{ y \cdot log\left(\frac{\mu}{1-\mu}\right) + log(1-\mu) \right\} = \\
= \left| \theta = log(\tfrac{\mu}{1-\mu}); \quad \mu = \frac{e^\theta}{1+e^\theta} \right| = \\
= exp\left\{ y \cdot \theta + log(\tfrac{1}{1+e^\theta}) \right\} =
\end{gathered}
\tag{5}
$$

Proof.

$$= exp\left\{y \cdot \theta - log(1 + e^{\theta})\right\}$$

gdzie:

$$\begin{cases} \theta := log\left(\frac{\mu}{1-\mu}\right) \\ \phi := 1 \\ b(\theta) := log(1 + e^{\theta}) \\ c(y, \phi) := 0 \end{cases}$$

Zatem rozkład binarny należy do rodziny wykładniczej. $\square$

Binarnej regresji logistycznej używamy do budowania modelu w przypadku gdy rozkład zmiennej zależnej Y jest określony funkcja rozkładu prawdopodobieństwa w nastepujacy sposób:

$$P(Y=1)=\mu \qquad \text{oraz} \qquad P(Y=0)=1-\mu$$

Wartość oczekiwana zmiennej Y to $E(Y)=\mu$. Model ten służy do przewidywania prawdopodobieństwa a posteriori[1] μ wystapienia sukcesu na podstawie danych (zmiennych niezależnych), korzystamy z niego w *Przykładzie 1*. Model regresji logistycznej, dla zmiennej objaśnianej o rozkładzie binarnym, dzieki użyciu funkcji łaczacej, zwraca wynik interpretowalny na całej przestrzeni liczb rzeczywistych. Wartość μ może zmieniać sie wraz ze zmiana wartości x, zatem zastepujemy μ przez $\mu(x)$, gdy chcemy opisać zależność od tej wartości. Dla ustalenia uwagi przyjmijmy logit za funkcje łaczaca:

$$g(\mu(x)) = logit(\mu(x)) = \beta_0 + \beta x \tag{6}$$

gdzie: β_0, β – współczynniki modelu.

[1] a posteriori – w filozofii termin oznaczajacy: po fakcie" lub w nastepstwie faktu"

Wyznaczajac funkcje odwrotna do g obliczamy, dla znanego x, prawdopodobieństwo a posteriori sukcesu:

$$g^{-1}(x) = \mu(x) = \frac{exp(\beta_0 + \beta x)}{1 + exp(\beta_0 + \beta x)} \tag{7}$$

Warto zauważyć, że:

$$szansa(x) = \frac{\mu(x)}{1 - \mu(x)} = e^{\beta_0 + \beta x}$$

Zatem:

$$\frac{szansa(x+1)}{szansa(x)} = \frac{e^{\beta_0 + \beta(x+1)}}{e^{\beta_0 + \beta x}} = e^{\beta}$$

Wiec szansa wzrasta e^{β} razy przy wzroście wartości x o 1. Wówczas $\mu(x)$ zazwyczaj rośnie badź maleje w sposób ciagły wraz ze wzrostem x. Jej monotoniczność zależy od znaku współczynnika β. Relacja ta została przedstawiona na rysunku przedstawionym na wykładzie.

Jeśli istnieje wiele zmiennych objaśniajacych, równość (6) rozszerzamy do postaci:

$$logit(\mu(x_1, x_2, ..., x_k)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$$

Analiza różnic w modelach

Rozważmy różnice miedzy analizowaniem wyniku regresji liniowej, a regresji logistycznej. Przykład omówimy na tablicy.

Przykład

Rozpoczniemy do aplikacyjnego charakteru uogólnionych modeli liniowych (GLM). W tabeli 1 przedstawiamy dane z 23 lotów kosmicznych przed katastrofa misji Challenger w 1986r. Tabela przedstawia temperature (F°) w czasie startu i informacje czy przynajmniej jedna uszczelka O-Rings rozszczelniła sie. Rozważmy model pierwszy z funkcja wiażaca $g(\mu) = \mu$ oraz model drugi z funkcja wiażaca $g(\mu) = logit(\mu) = log(\frac{\mu}{1-\mu})$.

Lp.	Temperatura	Usterka	Lp.	Temperatura	Usterka
1	53	1	13	70	0
2	57	1	14	70	1
3	58	1	15	72	0
4	63	1	16	73	0
5	66	0	17	75	0
6	67	0	18	75	1
7	67	0	19	76	0
8	67	0	20	76	0
9	68	0	21	78	0
10	69	0	22	79	0
11	70	0	23	81	0
12	70	0			

Table: Dane z ksiażki: Alan Agresti *An Introduction to Categorical Data Analysis*, Second Edition (tabela 4.10)

Uwaga: Usterka (1=wystapiła, 0=nie wystapiła)

Model pierwszy:

$$E(Usterka) = 2,888889 + temp \cdot (-0,037778)$$

Model drugi:

$$E(Usterka) = \frac{e^{16,798079 + temp \cdot (-0,263060)}}{1 + e^{16,798079 + temp \cdot (-0,263060)}}$$

Lp.	Model 1	Model 2	Lp.	Model 1	Model 2
1	0,8866666667	0,9456234303	13	0,2444444444	0,1657427394
2	0,7355555556	0,8585953369	14	0,2444444444	0,1657427394
3	0,6977777778	0,8235537071	15	0,1688888889	0,1050600495
4	0,5088888889	0,5560912516	16	0,1311111111	0,0827706285
5	0,3955555556	0,3626534324	17	0,0555555556	0,0506228136
6	0,3577777778	0,3042955086	18	0,0555555556	0,0506228136
7	0,3577777778	0,3042955086	19	0,0177777778	0,0393745994
8	0,3577777778	0,3042955086	20	0,0177777778	0,0393745994
9	0,32	0,2516210163	21	-0,057777778	0,0236471108
10	0,2822222222	0,2053729601	22	-0,095555556	0,0182774098
11	0,2444444444	0,1657427394	23	-0,17111111	0,0108813664
12	0,2444444444	0,1657427394			

Table: Wyestymowane prawdopodobieństwo dla dwóch modeli

Jak widać w tabeli 2, niektóre z naszych wyników dla modelu pierwszego nie należa do oczywistego przedziału prawdopodobieństwa [0;1], co niestety stawia pod znakiem zapytania ich interpretowalność. Taki problem nie wystepuje przy zastosowaniu logitowej funkcji wiażacej. Na pytanie dlaczego, odpowiemy przy okazji omawiania tej funkcji.

Frame Title

Frame Title

Frame Title

Frame Title

Frame Title

Frame Title

Frame Title

Frame Title

Frame Title

Frame Title

Frame Title