Spis treści

1 Wstęp

Celem tych notatek jest zwięzłe przedstawienie kompletu zagadnień związanych z szeroko pojętym uczeniem głębokim jako podejściem do Sztucznej Inteligencji (SI). Zaczynamy od minimalnego zbioru wymaganych tematów z zakresu rachunku prawdopodobieństwa i statystyki matematycznej. Następnie opisujemy podstawowe metody uczenia maszynowego z probabilistycznego punktu widzenia. W końcu przechodzimy do zasadniczej części związanej z uczeniem głębokim i sieciami neuronowymi. W każdej części staramy się przedstawiać opisywane tematy w sposób minimalistyczny, skupiając się głównie na matematycznej i ideowej, a nie implementacyjnej stronie zagadnień. Liczymy, iż takie podejście zapewni odpowiednio głębokie zrozumienie tematu, dzięki któremu dalsze studiowanie całej gamy specyficznych technicznych tematów nie sprawi żadnego problemu.

1.1 Notacja

W dalszej części tekstu będziemy stosować przedstawioną tutaj pokrótce notację. Wektory, które traktujemy jako elementy przestrzeni $\mathbb{R}^d$ ze standardowo zdefiniowanymi operacjami dodawania i mnożenia przez skalar będziemy oznaczali wytłuszczonymi małymi literami np. $\mathbf{x}, \mathbf{w}, \boldsymbol{\phi}$. Wielkość $\mathbf{x}_i$ będzie oznaczać dany element wektora (w tym przypadku i–ty element $\mathbf{x}$). Wielkość $\mathbf{x}^\mu$ będzie oznaczać pewien (w tym przypadku μ–ty) element pewnego zbioru wektorów. Macierze oraz wielowymiarowe tablice (zwane również niefortunnie tensorami) będziemy oznaczać wytłuszczonymi wielkimi literami np. $\mathbf{X}, \mathbf{W}, \boldsymbol{\Phi}$. Analogicznie jak w przypadku wektorów przez $\mathbf{X}_{i_1 i_2 \ldots i_k}$ będziemy oznaczać $(i_1, i_2, \ldots, i_k)$ element k–wymiarowej tablicy $\mathbf{X}$, natomiast $\mathbf{X}^\mu$ będzie oznaczać μ–ty element pewnego zbioru tablic.

1.2 Uczenie nadzorowane

Uczenie nadzorowane jest jednym z dwóch podstawowych (pomijając tzw. uczenie ze wzmocnieniem) paradygmatów w uczeniu maszynowym, którego ogólną ideą jest zdefiniowanie pewnego modelu odwzorowującego dane wejściowe na wyjściowe predykcje. Zakładamy w nim, iż mamy dostępny zbiór obserwacji w postaci uporządkowanych par $\mathcal{X} = \{(\mathbf{x}^\mu, y_\mu)\}_{\mu=1}^n$, gdzie $\mathbf{x} \in \mathbb{R}^d$ nazywamy wektorem cech a y jest prawidłową wartością odpowiedzi dla tych cech. Dwa najbardziej podstawowe przypadki zagadnień tego rodzaju to regresja oraz klasyfikacja. W przypadku regresji zmienna y przyjmuje wartości z przedziału liczb rzeczywistych. W przypadku klasyfikacji zmienna y przyjmuje wartości ze skończonego zbioru kategorii, przy czym wartości z tego zbioru nie powinny posiadać naturalnej tj. wynikającej z natury problemu, relacji porządku; gdy tak nie jest mamy do czynienia z regresją/klasyfikacją porządkową (z ang. *ordinal regression/classification*).

W jaki sposób tworzymy wspomniany model odwzorowujący $\mathbf{x}$ na y? W dalszych paragrafach poznamy różne metody, ale najczęściej (nie wchodząc teraz

w modelowanie probabilistyczne) modelem jest pewna rodzina funkcji postaci $\hat{y}(\mathbf{x}; \mathbf{w})$ parametryzowana skończoną liczbą parametrów, które możemy łącznie zapisać jako pewien wektor $\mathbf{w} \in \mathbb{R}^m$. Aby znaleźć parametry $\mathbf{w}$, dzięki którym dla konkretnego zagadnienia model będzie zadowalająco odwzorowywał cechy na predykcje (innymi słowy aby nauczyć model) wprowadzamy dodatkowo funkcjonał kosztu (z ang. *loss function*) $L[\hat{y}(\mathbf{x}; \mathbf{w}), y]$, który kwantyfikuje odpowiedzi modelu ϕ w stosunku do znanych prawidłowych odpowiedzi y. Trening modelu polega wówczas na znalezieniu parametrów $\mathbf{w}^*$, które minimalizują sumę wartości funkcji kosztu dla przykładów w zbiorze treningowym $\mathcal{X}$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{\mu=1}^{n} L[\hat{y}(\mathbf{x}^\mu; \mathbf{w}), y_\mu] \,. \tag{1.1}$$

Zauważmy, że takie podejście ma jedną zasadniczą wadę – istotnie nie interesuje nas tak naprawdę, jak model radzi sobie na zbiorze treningowym, tylko jak będzie radził sobie na nowych, niewidzianych wcześniej danych. Sytuację, w której model bardzo dobrze modeluje dane w zbiorze treningowym, ale słabo radzi sobie na nowych danych nazywamy przeuczeniem lub nadmiernym dopasowaniem (z ang. *overfitting*). Sytuację, w której model słabo radzi sobie zarówno na zbiorze treningowym, jak i na nowych danych nazywamy niedouczeniem lub niedopasowaniem (z ang. *underfitting*). Występowanie overfittingu i underfittingu jest powiązane z pojemnością (z ang. *capacity*) modelu. Złożony model o dużej pojemności potrafi dopasować się do bardzo skomplikowanych obserwacji, ale istnieje ryzyko jego przeuczenia (mówimy wówczas o *high variance*). Dla prostego modelu o małej pojemności istnieje z kolei ryzyko, iż nie ma on wystarczająco ekspresywności (mówimy wówczas o *high bias*).

1.3 Uczenie nienadzorowane

W przypadku uczenia nienadzorowanego naszym celem nie jest znalezienie modelu odwzorowującego cechy na predykcje. Chcemy raczej zrozumieć wewnętrzną strukturę danych. Modele tego rodzaju znajdują zastosowanie w analizie biznesowej, gdzie pozwalają, chociażby na analizę ważności poszczególnych wskaźników, czy wizualizację wysoko-wymiarowych danych. Nas będą interesować natomiast szczególnie generatywne modele nienadzorowane, za pomocą których modelujemy rozkład prawdopodobieństwa danych, czy to wprost poprzez funkcję gęstości prawdopodobieństwa, czy też jedynie jako model, z którego możemy próbkować nowe przykłady.

1.4 Praktyka uczenia maszynowego

W dalszej części nie będziemy skupiać się na detalach implementacyjnych. W naszej ocenie jednak dziedzina uczenia maszynowego jest przede wszystkim „nauką eksperymentalną" dlatego bardzo ważne jest empiryczne sprawdzenie różnych metod przed wybraniem ostatecznego modelu dla danego zagadnienia. W tym paragrafie opisujemy standardowe praktyki stosowane przy treningu modeli

uczenia maszynowego (głównie nadzorowanych operujących na danych tabelarycznych).

1.4.1 Przygotowanie danych

Kluczem do uzyskania dobrych wyników przy korzystaniu z algorytmów uczenia maszynowego jest odpowiednie przygotowanie danych (z ang. *preprocessing*). Typowo preprocessing składa się z:

- wczytania danych;
- eksploracji danych oraz wstępnego czyszczenia, w szczególności usunięcia jawnych wartości odstających (z ang. *outliers*) oraz cech posiadających zbyt dużo wartości brakujących;
- analizy rozkładu zmiennej docelowej oraz ewentualnej transformacji logarytmicznej, która poprawia stabilność numeryczną, gdy przewidywane wartości są dużymi dodatnimi liczbami rzeczywistymi, zmienia dziedzinę zmiennej objaśnianej z $\mathbb{R}_+$ na $\mathbb{R}$ oraz dodatkowo jest przykładem transformacji stabilizującej wariancję;
- **podziału zbioru na część treningową oraz testową**;
- dokonania skalowania i imputacji brakujących wartości cech (metody `.fit()` wywołujemy jedynie dla zbioru treningowego);
- usunięcia silnie skorelowanych cech;
- zakodowania wartości kategorycznych za pomocą tzw. *one–hot encoding* pamiętając o *dummy variable trap* – jedną z k kategorii kodujemy za pomocą wektora *one–hot* długości $n - 1$, aby uniknąć zależności liniowej między cechami (opcja `drop="first"` w `OneHotEncoder` w scikit-learn);
- wykonania feature engineering – dodania wielomianów cech do naszych danych lub skonstruowania innych cech (np. cech określających miesiąc, dzień itp.);

Podział zbioru na część treningową i testową jest najważniejszym etapem preprocessingu. Zbiór testowy wydzielamy, aby po wytrenowaniu modelu sprawdzić, jak poradzi on sobie na nowych, niewidzianych wcześniej danych. Powinniśmy go traktować jako dane, które będziemy w przyszłości dostawać po wdrożeniu modelu do realnego systemu. Takie dane również będziemy musieli przeskalować, zakodować itp., ale parametry potrzebne do wykonania tych transformacji możemy wziąć jedynie z dostępnego wcześniej zbioru treningowego. Wykorzystanie danych testowych w procesie treningu to **błąd wycieku danych** (z ang. *data leakage*). Skutkuje on niepoprawnym, nadmiernie optymistycznym oszacowaniem jakości modelu.

1.4.2 Tuning hiperparametrów i walidacja skrośna

Praktycznie wszystkie modele uczenia maszynowego mają hiperparametry, często liczne, które w zauważalny sposób wpływają na wyniki, a szczególnie na underfitting i overfitting. Ich wartości trzeba dobrać zatem dość dokładnie. Proces doboru hiperparametrów nazywa się tuningiem hiperparametrów (z ang. *hyperparameter tuning*).

Istnieje na to wiele sposobów. Większość z nich polega na tym, że trenuje się za każdym razem model z nowym zestawem hiperparametrów i wybiera się ten zestaw, który pozwala uzyskać najlepsze wyniki. Metody głównie różnią się między sobą sposobem doboru kandydujących zestawów hiperparametrów. Najprostsze i najpopularniejsze to:

- pełne przeszukiwanie (z ang. *grid search*) – definiujemy możliwe wartości dla różnych hiperparametrów, a metoda sprawdza ich wszystkie możliwe kombinacje (czyli siatkę),
- losowe przeszukiwanie (z ang. *randomized search*) – definiujemy możliwe wartości jak w pełnym przeszukiwaniu, ale sprawdzamy tylko ograniczoną liczbę losowo wybranych kombinacji.

Jak ocenić, jak dobry jest jakiś zestaw hiperparametrów? Nie możemy sprawdzić tego na zbiorze treningowym – wyniki byłyby zbyt optymistyczne. Nie możemy wykorzystać zbioru testowego – mielibyśmy wyciek danych, bo wybieralibyśmy model explicite pod nasz zbiór testowy. Trzeba zatem osobnego zbioru, na którym będziemy na bieżąco sprawdzać jakość modeli dla różnych hiperparametrów. Jest to zbiór walidacyjny (z ang. *validation set*). Zbiór taki wycina się ze zbioru treningowego.

Jednorazowy podział zbioru na części nazywa się *split validation* lub *holdout*. Używamy go, gdy mamy sporo danych, i 10-20% zbioru jako dane walidacyjne czy testowe to dość dużo, żeby mieć przyzwoite oszacowanie. Zbyt mały zbiór walidacyjny czy testowy da nam mało wiarygodne wyniki – nie da się nawet powiedzieć, czy zbyt pesymistyczne, czy optymistyczne. W praktyce niestety często mamy mało danych. Trzeba zatem jakiejś magicznej metody, która stworzy nam więcej zbiorów walidacyjnych z tej samej ilości danych. Taką metodą jest walidacja skrośna (z ang. *cross validation*, CV). Polega na tym, że dzielimy zbiór treningowy na K równych podzbiorów, tzw. foldów. Każdy podzbiór po kolei staje się zbiorem walidacyjnym, a pozostałe łączymy w zbiór treningowy. Trenujemy zatem K modeli dla tego samego zestawu hiperparametrów i każdy testujemy na zbiorze walidacyjnym. Mamy K wyników dla zbiorów walidacyjnych, które możemy uśrednić (i ewentualnie obliczyć odchylenie standardowe). Takie wyniki są znacznie bardziej wiarygodne.

2 Rachunek prawdopodobieństwa i statystyka matematyczna

2.1 Przestrzeń probabilistyczna

Pojęciem pierwotnym w rachunku prawdopodobieństwa jest pojęcie **przestrzeni zdarzeń elementarnych**, którą oznaczamy Ω. W przypadku doświadczeń losowych przestrzeń zdarzeń elementarnych jest zbiorem wszystkich niepodzielnych wyników obserwacji.

Definicja 2.1 (Rodziny zdarzeń)**.** *Niech Ω będzie przestrzenią zdarzeń elementarnych. Rodziną zdarzeń nazwiemy rodzinę zbiorów $\mathcal{F}$ taką, że*

1. $\Omega \in \mathcal{F}$.
2. *Jeśli* $A \in \mathcal{F}$ *to* $\Omega \setminus A \in \mathcal{F}$.
3. *Jeśli* $A_1, A_2, \ldots \in \mathcal{F}$ *to* $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Rodzinę zdarzeń $\mathcal{F}$ nazywamy σ–ciałem zdarzeń losowych.

Definicja 2.2 (Zdarzenia losowego)**.** *Zdarzeniem losowym nazywamy dowolny zbiór należący do rodziny zdarzeń $\mathcal{F}$. W szczególności Ω nazwiemy zdarzeniem pewnym, a $\varnothing$ – zdarzeniem niemożliwym.*

Definicja 2.3 (Rozkładu prawdopodobieństwa)**.** *Niech dana będzie przestrzeń zdarzeń elementarnych Ω i rodzina zdarzeń $\mathcal{F}$. Rozkładem prawdopodobieństwa nazwiemy funkcję $P : \mathcal{F} \mapsto [0; 1]$ spełniającą*

1. *Dla każdego $A \in \mathcal{F}$ zachodzi $P(A) \geq 0$.*
2. $P(\Omega) = 1$.
3. *Dla każdego ciągu zdarzeń parami rozłącznych $A_1, A_2, \ldots$, $\forall i \neq j : A_i \cap A_j = \varnothing$ zachodzi*
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)\,.$$

Twierdzenie 2.1. *Niech P będzie rozkładem prawdopodobieństwa w rodzinie zdarzeń $\mathcal{F}$, wówczas*

1. $P(\varnothing) = 1$.
2. *(Addytywność) Dla dowolnych zdarzeń $A_1, \ldots, A_n$ parami rozłącznych zachodzi*
$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)\,.$$
3. *Dla dowolnego zdarzenia A zachodzi*
$$P(A') = 1 - P(A)\,.$$

4. *Dla dowolnych zdarzeń A, B zachodzi*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)\,.$$

5. *Jeśli $A \subset B$ to $P(A) \leq P(B)$.*

6. *Dla każdego zdarzenia A zachodzi $P(A) \leq 1$.*

Definicja 2.4 (Przestrzeni probabilistycznej)**.** *Przestrzenią probabilistyczną nazywamy uporządkowaną trójkę $(\Omega, \mathcal{F}, P)$, gdzie Ω jest przestrzenią zdarzeń elementarnych, $\mathcal{F}$ – rodziną zdarzeń określoną na Ω, a P – rozkładem prawdopodobieństwa w $\mathcal{F}$.*

2.2 Prawdopodobieństwo warunkowe i niezależność zdarzeń

Definicja 2.5 (Prawdopodobieństwa warunkowego)**.** *Jeśli $P(B) > 0$ to prawdopodobieństwem $P(A \mid B)$ zdarzenia A pod warunkiem zdarzenia B nazywamy iloraz prawdopodobieństw*

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}\,.$$

Można pokazać, iż prawdopodobieństwo warunkowe jako funkcja zdarzenia A przy ustalonym zdarzeniu B spełnia wszystkie aksjomaty rozkładu prawdopodobieństwa.

Twierdzenie 2.2 (O wielokrotnym warunkowaniu)**.** *Dla dowolnych zdarzeń $A_1, \ldots, A_n$ takich, że $P(A_1, \ldots, A_n) > 0$ zachodzi*

$$\begin{aligned} &P(A_1, \ldots, A_n) = \\ &P(A_n \mid A_{n-1}, \ldots, A_1) \ldots P(A_2 \mid A_1) P(A_1)\,, \end{aligned}$$

gdzie dla uproszczenia zapisu używamy $P(A, B)$ w znaczeniu $P(A \cap B)$.

Definicja 2.6 (Niezależności pary zdarzeń)**.** *Zdarzenia A, B nazwiemy niezależnymi jeśli*

$$P(A, B) = P(A)P(B)\,.$$

Zauważmy, iż w przypadku $P(B) > 0$ powyższa definicja jest równoważna bardziej intuicyjnej wynikającej z definicji prawdopodobieństwa warunkowego, mianowicie zdarzenia A, B nazywamy niezależnymi jeśli $P(A \mid B) = P(A)$.

Definicja 2.7 (Niezależności zdarzeń)**.** *Zdarzenia $A_1, \ldots, A_n$ nazwiemy niezależnymi, jeśli dla dowolnych wskaźników $k_1, \ldots, k_s$, gdzie $1 \leq k_1 < \ldots < k_s \leq n$ zachodzi*

$$P(A_{k_1}, \ldots, A_{k_s}) = P(A_{k_1}) \ldots P(A_{k_s})\,.$$

Twierdzenie 2.3 (O łącznym prawdopodobieństwie niezależnych zdarzeń)**.** *Jeśli zdarzenia $A_1, \ldots, A_n$ są niezależne to*

$$P(A_1, \ldots, A_n) = P(A_1) \ldots P(A_n)\,.$$

Definicja 2.8 (Warunkowej niezależności pary zdarzeń)**.** *Zdarzenia A, B są warunkowo niezależne względem C jeśli*

$$P(A, B \mid C) = P(A \mid C)P(B \mid C)\,.$$

Definicja 2.9 (Warunkowej niezależności zdarzeń)**.** *Zdarzenia $A_1, \ldots, A_n$ są warunkowo niezależne względem B jeśli dla dowolnych wskaźników $k_1, \ldots, k_s$, gdzie $1 \leq k_1 < \ldots < k_s \leq n$ zachodzi*

$$P(A_{k_1}, \ldots, A_{k_s} \mid B) = P(A_{k_1} \mid B) \ldots P(A_{k_s} \mid B)\,.$$

Twierdzenie 2.4. *Jeśli zdarzenia $A_1, \ldots, A_n$ są warunkowo niezależne względem B to*

$$P(A_1, \ldots, A_n \mid B) = P(A_1 \mid B) \ldots P(A_n \mid B)\,.$$

2.3 Prawdopodobieństwo całkowite i wzór Bayesa

Definicja 2.10 (Układu zupełnego zdarzeń)**.** *Jeśli zdarzenia $A_1, A_2 \ldots,$ są parami rozłączne oraz $\bigcup_{i=1}^{\infty} A_i = \Omega$ to zbiór zdarzeń $\{A_i\}$ nazywamy układem zupełnym.*

Twierdzenie 2.5 (O prawdopodobieństwie całkowitym)**.** *Jeśli zdarzenia A_i (gdzie i przebiega przeliczalny zbiór wartości) tworzą układ zupełny zdarzeń oraz $P(A_i) > 0$ dla każdego i, to dla dowolnego zdarzenia B zachodzi*

$$P(B) = \sum_i P(B \mid A_i)P(A_i)\,.$$

Twierdzenie 2.6 (Bayesa)**.** *Jeśli zdarzenia A_i spełniają założenia tw. o prawdopodobieństwie całkowitym oraz $P(B) > 0$, to dla każdego zdarzenia A_j z rozpatrywanego układu zdarzeń zachodzi*

$$P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{P(B)} = \frac{P(B \mid A_j)P(A_j)}{\sum_i P(B \mid A_i)P(A_i)}\,.$$

*Prawdopodobieństwa $P(A_j)$ nazywamy prawdopodobieństwami **a priori**, a $P(A_j \mid B)$ – **a posteriori**. Prawdopodobieństwo $P(B \mid A_j)$ nazywamy **wiarygodnością**.*

2.4 Zmienne losowe

Definicja 2.11 (Zmiennej losowej)**.** *Niech $(\Omega, \mathcal{F}, P)$ będzie przestrzenią probabilistyczną. Niech dany będzie również drugi zbiór $\mathcal{X}$, w którym wyróżniamy σ–ciało $\mathcal{F}_{\mathcal{X}}$. Zmienną losową X nazywamy odwzorowanie*

$$X : \Omega \mapsto \mathcal{X}$$

takie, że dla każdego $A \in \mathcal{F}_{\mathcal{X}}$ *zachodzi warunek*

$$\{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}\,.$$

W szczególności jeśli $\mathcal{X} = \mathbb{R}$ to zmienną losową X nazywamy **zmienną losową rzeczywistą**, natomiast jeśli $\mathcal{X} = \mathbb{R}^n$ to zmienną losową $\mathbf{X}$ nazywamy **zmienną losową** n**–wymiarową**.

Definicja 2.12 (Rozkładu prawdopodobieństwa zmiennej losowej)**.** *Niech* $X : \Omega \mapsto \mathcal{X}$ *będzie zmienną losową. Funkcję* $P_X : \mathcal{F}_{\mathcal{X}} \mapsto [0;1]$ *określoną jako*

$$P_X(A) = P(\{\omega \in \Omega : X(\omega) \in A\})$$

nazywamy rozkładem prawdopodobieństwa zmiennej losowej X*.*

W dalszym ciągu będziemy stosować notację uproszczoną pomijając indeks dolny X tj.

$$P(X \in A) := P(\{\omega \in \Omega : X(\omega) \in A\})\,.$$

Znajomość rozkładu zmiennej losowej X pozwala badać własności tej zmiennej bez znajomości przestrzeni probabilistycznej $(\Omega, \mathcal{F}, P)$, na której owa zmienna jest określona. Możemy mianowicie zawsze rozpatrywać tę zmienną jako określoną na przestrzeni probabilistycznej $(\mathcal{X}, \mathcal{F}_{\mathcal{X}}, P_X)$.

Definicja 2.13 (Dystrybuanty zmiennej losowej rzeczywistej)**.** *Niech* $X : \Omega \mapsto \mathbb{R}$ *będzie zmienną losową rzeczywistą. Dystrybuantą zmiennej losowej* X *nazywamy funkcję* $F : \mathbb{R} \mapsto [0;1]$ *zdefiniowaną jako*

$$F(x) = P(X \leq x)\,.$$

Twierdzenie 2.7 (Własności dystrybuanty zmiennej rzeczywistej)**.** *Niech* F *będzie dystrybuantą zmiennej losowej rzeczywistej* X*, wówczas*

1. *Jeśli* $a < b$ *to* $F(b) - F(a) = P(X \in (a;b])$*.*
2. *Funkcja* F *jest niemalejąca.*
3. $\lim_{x\to-\infty} F(x) = 0$ *oraz* $\lim_{x\to+\infty} F(x) = 1$*.*
4. F *jest prawostronnie ciągła.*
5. F *jest ciągła w* x_0 *wtedy i tylko wtedy, gdy* $P(X = x_0) = 0$*.*

Definicja 2.14 (Rozkładu dyskretnego zmiennej losowej rzeczywistej)**.** *Mówimy, że zmienna losowa rzeczywista* X *ma rozkład dyskretny jeśli istnieje skończony lub przeliczalny zbiór* $\mathcal{S} \subset \mathbb{R}$ *taki, że*

$$P(X \in \mathcal{S}) = 1\,.$$

Dla takiej zmiennej określa się funkcję prawdopodobieństwa (z ang. probability mass function, pmf)

$$p(x) = P(X = x)\,, \quad x \in \mathcal{S}\,.$$

Zauważmy, że jeśli zmienna X ma rozkład dyskretny to zbiór $\mathcal{S}$ ma postać $\{x_1, \ldots, x_k\}$ dla pewnych $x_i \in \mathbb{R}$.

Definicja 2.15 (Rozkładu ciągłego zmiennej losowej rzeczywistej)**.** *Mówimy, że zmienna losowa rzeczywista X ma rozkład ciągły jeśli istnieje funkcja $p : \mathbb{R} \mapsto [0; +\infty)$ taka, że dla dowolnego przedziału $(a; b)$ zachodzi*

$$P(X \in (a; b)) = \int_a^b p(x)\,\mathrm{d}x \ .$$

Funkcję p nazywamy gęstością rozkładu prawdopodobieństwa (z ang. probability density function, pdf).

Jeśli X jest zmienną losową rzeczywistą o rozkładzie ciągłym to wartość dystrybuanty jest dana przez

$$F(x) = \int_{-\infty}^{x} p(x')\,\mathrm{d}x' \ . \tag{2.1}$$

W szczególności dystrybuanta F jest ciągła oraz dla każdego x zachodzi $P(X = x) = 0$.

Definicja 2.16 (Wartości oczekiwanej zmiennej losowej rzeczywistej)**.** *Wartością oczekiwaną zmiennej losowej rzeczywistej X nazywamy liczbę m określoną wzorem*

$$m = \int_{-\infty}^{+\infty} xp(x)\,\mathrm{d}x$$

dla rozkładu ciągłego oraz

$$m = \sum_{x \in \mathcal{S}} xp(x)$$

dla rozkładu dyskretnego. Stosujemy dodatkowo oznaczenie $\mathbb{E}[X] := m$.

Definicja 2.17 (Wariancji zmiennej losowej rzeczywistej)**.** *Wariancją zmiennej losowej rzeczywistej X nazywamy liczbę σ określoną wzorem*

$$\sigma^2 = \mathbb{E}[(X - m)^2]\,,$$

*gdzie $m = \mathbb{E}[X]$. Stosujemy dodatkowo oznaczenie $\mathbb{V}[X] := \sigma^2$. **Odchyleniem standardowym** zmiennej losowej X nazywamy pierwiastek jej wariancji*

$$\sigma = \sqrt{\mathbb{V}[X]}\,.$$

Twierdzenie 2.8 (Własności wariancji)**.** *Niech X będzie zmienną losową rzeczywistą, wówczas*

1. $\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

2. *Jeśli zmienna X ma skończoną wariancję, to dla dowolnych $a, b \in \mathbb{R}$ zachodzi*
$$\mathbb{V}[aX + b] = a^2\mathbb{V}[X]\,.$$

3. $\mathbb{V}[X] = 0$ *wtedy i tylko wtedy, gdy istnieje $x_0 \in \mathbb{R}$ takie, że*
$$P(X \neq x_0) = 0\,.$$

Definicja 2.18 (Zmiennej standaryzowanej)**.** *Zmienną losową o wartości oczekiwanej 0 i wariancji 1 nazywamy zmienną standaryzowaną. Jeśli X jest dowolną zmienną o niezerowej wariancji, to*
$$Z := \frac{X - m}{\sigma}$$
jest zmienną standaryzowaną, ponieważ
$$\mathbb{E}[Z] = \frac{1}{\sigma}\mathbb{E}[X - m] = \frac{1}{\sigma}(m - m) = 0$$
oraz
$$\mathbb{V}[Z] = \frac{1}{\sigma^2}\mathbb{V}[X] = \frac{\sigma^2}{\sigma^2} = 1\,.$$

Twierdzenie 2.9 (Nierówność Czebyszewa)**.** *Jeśli zmienna losowa X ma skończoną wartość średnią m i wariancję σ^2, to dla dowolnego $\epsilon > 0$ zachodzi*
$$P(|X - m| \geq \epsilon\sigma) \leq \frac{1}{\epsilon^2}\,.$$

Definicja 2.19 (Kwantyla)**.** *Kwantylem rzędu $p \in (0; 1)$ zmiennej losowej o dystrybuancie F nazywamy dowolną liczbę q_p taką, że*
$$F(q_p^-) \leq p \leq F(q_p)\,.$$

Kwantyl rzędu 0.5 nazywamy ***medianą****, kwantyl rzędu 0.25 –* ***dolnym kwartylem****, a kwantyl rzędu 0.75 –* ***górnym kwartylem****. Jeśli X ma rozkład ciągły, to kwantylem rzędu p jest dowolne rozwiązanie równania*
$$F(q_p) = p\,.$$

Definicja 2.20 (Mody)**.** *Modą zmiennej losowej o rozkładzie dyskretnym nazywa się dowolne maksimum funkcji prawdopodobieństwa tego rozkładu.*

Modą zmiennej losowej o rozkładzie ciągłym nazywa się dowolne maksimum lokalne gęstości tego rozkładu.

2.5 Ważne rozkłady jednowymiarowe

Definicja 2.21 (Rozkładu jednopunktowego)**.** *Jeśli X jest zmienną losową rzeczywistą o rozkładzie dyskretnym i $\mathcal{S} = \{x_0\}$, to mówimy, że X ma rozkład jednopunktowy, wówczas*
$$m = x_0, \quad \sigma^2 = 0\,.$$

Definicja 2.22 (Rozkładu dwupunktowego)**.** *Jeśli X jest zmienną losową rzeczywistą o rozkładzie dyskretnym i $\mathcal{S} = \{x_1, x_2\}$ oraz $p(x_1) = p$, to mówimy, że X ma rozkład dwupunktowy z parametrem p, wówczas*

$$m = x_1 p + x_2(1-p)\,, \quad \sigma^2 = p(1-p)(x_1 - x_2)^2\,.$$

Jeśli $x_1 = 1$ i $x_2 = 0$ to taki rozkład dwupunktowy nazywamy ***rozkładem zero–jedynkowym*** *(lub rozkładem Bernoulliego) i oznaczamy jako $X \sim \mathrm{Ber}(p)$.*

Definicja 2.23 (Schematu dwumianowego)**.** *Rozważmy doświadczenie losowe o dwu możliwych wynikach: sukces osiągamy z prawdopodobieństwem p, porażkę z prawdopodobieństwem $1-p$. Doświadczenie tego rodzaju nazywamy* ***próbą Bernoulliego****. Doświadczenie takie jest modelowane zmienną losową o rozkładzie dwupunktowym z parametrem p. Schematem dwumianowym (lub schematem Bernoulliego) nazywamy doświadczenie polegające na n–krotnym powtórzeniu próby Bernoulliego, przy założeniu, iż poszczególne próby są od siebie niezależne.*

Definicja 2.24 (Rozkładu dwumianowego)**.** *Niech X będzie zmienną losową taką, że X jest liczbą sukcesów w schemacie dwumianowym długości n z prawdopodobieństwem sukcesu w każdej próbie równym p. Wówczas*

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}\,.$$

Rozkład prawdopodobieństwa określony powyższym wzorem nazywam się rozkładem dwumianowym o parametrach n, p. Jeśli zmienna X ma rozkład dwumianowy to stosujemy notację $X \sim \mathrm{Bin}(n,p)$. Jeśli $X \sim \mathrm{Bin}(n,p)$ to

$$m = np\,, \quad \sigma^2 = np(1-p)\,.$$

Definicja 2.25 (Rozkładu geometrycznego)**.** *Mówimy, że zmienna losowa X ma rozkład geometryczny z parametrem $p \in (0;1)$, tj. $X \sim \mathrm{Geo}(p)$, jeśli $\mathcal{S} = \mathbb{N} \setminus \{0\}$, a funkcja prawdopodobieństwa ma postać*

$$p(x) = (1-p)^{x-1} p\,.$$

Zmienna X opisuje czas oczekiwania na pierwszy sukces w schemacie dwumianowym o nieskończonej długości. Jeśli $X \sim \mathrm{Geo}(p)$, to

$$m = p^{-1}\,, \quad \sigma^2 = \frac{1-p}{p^2}\,.$$

Definicja 2.26 (Rozkładu Poissona)**.** *Jeśli zmienna X o wartościach w $\mathbb{N}$ opisuje liczbę wystąpień pewnego powtarzalnego zdarzenia w przedziale czasowym $[0;t]$, przy czym spełnione są następujące założenia:*

- *powtórzenia zdarzenia występują niezależnie od siebie;*
- *„intensywność" wystąpień r jest stała;*

- *w danej chwili (rozumianej jako odpowiednio mały przedział) może zajść co najwyżej jedno zdarzenie*

to zmienna ta ma rozkład Poissona z parametrem $\lambda = rt$, tj. $X \sim \mathrm{Pos}(\lambda)$. Jeśli $X \sim \mathrm{Pos}(\lambda)$, to

$$P(X = k) = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}\,.$$

Ponadto

$$m = \lambda\,, \quad \sigma^2 = \lambda\,.$$

Twierdzenie 2.10 (Poissona)**.** *Niech (X_n) będzie ciągiem zmiennych losowych takich, że $X_n \sim \mathrm{Bin}(n, p_n)$, gdzie (p_n) jest ciągiem takim, że*

$$\lim_{n\to\infty} np_n = \lambda$$

dla pewnej liczby $\lambda > 0$. Wówczas

$$\lim_{n\to\infty} P(X_n = k) = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}\,.$$

Definicja 2.27 (Rozkładu jednostajnego)**.** *Mówimy, że zmienna X o rozkładzie ciągłym ma rozkład jednostajny na przedziale $[a; b]$ tzn. $X \sim \mathcal{U}(a, b)$ jeśli jej gęstość wyraża się wzorem*

$$p(x) = \begin{cases} \frac{1}{b-a}\,, & x \in [a; b] \\ 0\,, & x \notin [a; b] \end{cases}\,.$$

Jeśli $X \sim \mathcal{U}(a, b)$, to

$$m = \frac{a+b}{2}\,, \quad \sigma^2 = \frac{(b-a)^2}{12}\,.$$

Definicja 2.28 (Rozkładu wykładniczego)**.** *Niech T będzie zmienną modelującą czas oczekiwania na pierwsze zdarzenie w ciągu zdarzeń takim, że czas wystąpienia każdego z nich w przedziale $[0; t]$ jest opisany przez zmienną $X \sim \mathrm{Pos}(\lambda t)$. wtedy*

$$P(T > t) = P(X = 0) = \mathrm{e}^{-\lambda t}$$

oraz

$$P(T > 0) = 1\,.$$

Mówimy wtedy, że T ma rozkład wykładniczy z parametrem λ, tzn. $T \sim \mathrm{Exp}(\lambda)$. Gęstość rozkładu wykładniczego ma postać

$$p(t) = \begin{cases} 0\,, & t \le 0 \\ \lambda\mathrm{e}^{-\lambda t}\,, & t > 0 \end{cases}\,.$$

Jeśli $T \sim \mathrm{Exp}(\lambda)$, to

$$m = \lambda^{-1}\,, \quad \sigma^2 = \lambda^{-2}\,.$$

Definicja 2.29 (Rozkładu normalnego)**.** *Mówimy, że zmienna losowa X o gęstości $\phi(x;\mu,\sigma^2)$ ma rozkład normalny z parametrami $\mu \in \mathbb{R}, \sigma^2 \in [0;+\infty)$, tzn. $X \sim \mathcal{N}(\mu,\sigma^2)$, jeśli*

$$\phi(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) .$$

Jeśli $X \sim \mathcal{N}(\mu,\sigma^2)$, to

$$m = \mu\,, \quad \mathbb{V}[X] = \sigma^2\,.$$

Jeśli $X \sim \mathcal{N}(\mu,\sigma^2)$, to

$$\begin{aligned} P(|X-m| > 2\sigma) &\approx 0.0455003\,, \\ P(|X-m| > 3\sigma) &\approx 0.0026998\,. \end{aligned} \tag{2.2}$$

Obserwacje te nazywamy odpowiednio regułą 5% oraz regułą 3σ. Rozkład $\mathcal{N}(0,1)$ nazywamy **standardowym rozkładem normalnym**. Jeśli $X \sim \mathcal{N}(\mu,\sigma^2)$, to zmienna standaryzowana $Z = (X-\mu)/\sigma$ ma standardowy rozkład normalny.

2.6 Zmienne losowe wielowymiarowe

Definicja 2.30 (Dystrybuanty zmiennej losowej wielowymiarowej)**.** *Niech $\mathbf{X} : \Omega \mapsto \mathbb{R}^n$ będzie zmienną losową n–wymiarową. Dystrybuantą zmiennej losowej $\mathbf{X}$ nazywamy funkcję $F : \mathbb{R}^n \mapsto [0;1]$ zdefiniowaną jako*

$$F(\mathbf{x}) = P(\mathbf{X}_1 \leq \mathbf{x}_1, \ldots, \mathbf{X}_n \leq \mathbf{x}_n)$$

Definicja 2.31 (Rozkładu dyskretnego zmiennej losowej wielowymiarowej)**.** *Mówimy, że n–wymiarowa zmienna losowa $\mathbf{X}$ ma rozkład dyskretny jeśli istnieje zbiór skończony lub przeliczalny $\mathcal{S} \subset \mathbb{R}^n$ taki, że*

$$P(\mathbf{X} \in \mathcal{S}) = 1\,.$$

Dla takiej zmiennej określa się funkcję prawdopodobieństwa (z ang. probability mass function, pmf)

$$p(\mathbf{x}) = P(\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_n = \mathbf{x}_n)\,, \quad \mathbf{x} \in \mathcal{S}\,.$$

Zauważmy, że jeśli zmienna $\mathbf{X}$ ma rozkład dyskretny to zbiór $\mathcal{S}$ ma postać $\{\mathbf{x}^1, \ldots, \mathbf{x}^k\}$ dla pewnych $\mathbf{x}^i \in \mathbb{R}$.

Definicja 2.32 (Rozkładu ciągłego zmiennej losowej wielowymiarowej)**.** *Mówimy, że n–wymiarowa zmienna losowa $\mathbf{X}$ ma rozkład ciągły jeśli istnieje funkcja $p : \mathbb{R}^n \mapsto [0;+\infty)$ taka, że dla dowolnych przedziałów $(a_i;b_i)$ zachodzi*

$$P(\mathbf{X}_1 \in (a_1;b_1), \ldots, \mathbf{X}_n \in (a_n;b_n)) = \int_{a_1}^{b_1} \ldots \int_{a_n}^{b_n} p(\mathbf{x})\,\mathrm{d}^n\mathbf{x}\,.$$

Funkcję p nazywamy gęstością rozkładu prawdopodobieństwa (z ang. probability density function, pdf).

Zauważmy, że jeśli $\mathbf{X}$ jest zmienną losową wielowymiarową o rozkładzie ciągłym to wartość dystrybuanty jest związana z gęstością rozkładu prawdopodobieństwa poprzez

$$F(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}_1} \dots \int_{-\infty}^{\mathbf{x}_n} p(\mathbf{x}') \, \mathrm{d}^n \mathbf{x}' \ . \tag{2.3}$$

Definicja 2.33 (Wartości oczekiwanej zmiennej losowej wielowymiarowej)**.** *Wartością oczekiwaną zmiennej losowej wielowymiarowej nazywamy n–elementowy wektor rzeczywisty $\mathbf{m}$, którego elementy są określone wzorem*

$$\mathbf{m} = \int_{\mathbb{R}^n} \mathbf{x} p(\mathbf{x}) \, \mathrm{d}^n \mathbf{x}$$

dla rozkładu ciągłego oraz

$$\mathbf{m} = \sum_{\mathbf{x} \in \mathcal{S}} \mathbf{x} p(\mathbf{x})$$

dla rozkładu dyskretnego. Stosujemy dodatkowo oznaczenie $\mathbb{E}[\mathbf{X}] := \mathbf{m}$.

Definicja 2.34 (Rozkładu brzegowego)**.** *Niech $\mathbf{X}$ będzie n–wymiarową zmienną losową. Weźmy dwa zbiory indeksów $\{i_1, \dots, i_k\} \subset \{1, \dots, n\}$ i $\{j_1, \dots, j_{n-k}\} = \{1, \dots, n\} \setminus \{i_1, \dots, i_k\}$. $(n-k)$–wymiarowym rozkładem brzegowym zmiennej $\mathbf{X}$ (względem zmiennych $j_1, \dots, j_{n-k}$) nazywamy rozkład prawdopodobieństwa na przestrzeni $\mathbb{R}^{n-k}$ określony wzorem*

$$\begin{aligned} & P_{\mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_{n-k}}}(A) \\ & = P\left(\left[\mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_{n-k}} \right] \in A \right) \\ & = P\left(\left[\mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_{n-k}} \right] \in A, \mathbf{X}_{i_1} \in \mathbb{R}, \dots, \mathbf{X}_{i_k} \in \mathbb{R} \right) \end{aligned}$$

Niech $\mathbf{X}$ będzie n–wymiarową zmienną losową o dystrybuancie F. Dystrybuanta rozkładu brzegowego $\mathbf{X}$ względem zmiennych $\mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_{n-k}}$ spełnia równość

$$F_{\mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_{n-k}}}(\mathbf{x}_{j_1}, \dots, \mathbf{x}_{j_{n-k}}) = \lim_{\mathbf{x}_{i_1} \to \infty, \dots, \mathbf{x}_{i_k} \to \infty} F(\mathbf{x}_1, \dots, \mathbf{x}_n) \, . \tag{2.4}$$

Dystrybuanta ta nosi nazwę $(n-k)$–wymiarowej dystrybuanty brzegowej.

Jeśli $\mathbf{X}$ jest n–wymiarową zmienną losową o rozkładzie ciągłym, to rozkłady brzegowe są także rozkładami ciągłymi o gęstościach

$$p_{\mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_{n-k}}}(\mathbf{x}_{j_1}, \dots, \mathbf{x}_{j_{n-k}}) = \int_{\mathbb{R}^k} p(\mathbf{x}') \, \mathrm{d}\mathbf{x}_{i_1} \dots \mathrm{d}\mathbf{x}_{i_k} \ . \tag{2.5}$$

Jeśli z kolei $\mathbf{X}$ ma rozkład dyskretny, to rozkłady brzegowe także są rozkładami dyskretnymi z funkcjami prawdopodobieństwa

$$p_{\mathbf{X}_{j_1}, \dots, \mathbf{X}_{j_{n-k}}}(\mathbf{x}_{j_1}, \dots, \mathbf{x}_{j_{n-k}}) = \sum_{i_1, \dots, i_k} p(\mathbf{x}) \, . \tag{2.6}$$

Twierdzenie 2.11. *Niech zmienna n–wymiarowa $\mathbf{X}$ ma rozkład ciągły o gęstości $p_{\mathbf{X}}$ i niech $\mathbf{Y}_i = \boldsymbol{\varphi}_i(\mathbf{X})$ dla $i = 1, \ldots, n$. Jeśli odwzorowanie $\boldsymbol{\varphi}$ jest różniczkowalne i odwracalne, przy czym odwzorowanie odwrotne $\boldsymbol{\psi} = \boldsymbol{\varphi}^{-1}$ jest różniczkowalne, to n–wymiarowa zmienna $\mathbf{Y}$ ma rozkład o gęstości*

$$p_{\mathbf{Y}}(\mathbf{y}) = |J| p_{\mathbf{X}}(\boldsymbol{\psi}(\mathbf{y}))\,,$$

gdzie $J := \det\left[\frac{\partial \boldsymbol{\psi}_j}{\partial \mathbf{y}_i}\right]$ jest jakobianem odwzorowania $\boldsymbol{\psi}$.

2.7 Niezależność zmiennych losowych

Definicja 2.35 (Niezależności zmiennych losowych)**.** *Niech $X_i : \Omega \mapsto \mathcal{X}$ będą zmiennymi losowymi. Zmienne $X_1, \ldots, X_n$ nazywamy niezależnymi jeżeli dla dowolnych $A_1, \ldots, A_n \in \mathcal{F}_{\mathcal{X}}$ zachodzi równość*

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \ldots P(X_n \in A_n)\,.$$

Twierdzenie 2.12. *Zmienne losowe wielowymiarowe $\mathbf{X}, \mathbf{Y}$ o rozkładzie dyskretnym lub ciągłym z funkcją lub gęstością prawdopodobieństwa $p(\mathbf{x}, \mathbf{y})$ są niezależne wtedy i tylko wtedy, gdy dla dowolnych $\mathbf{x}, \mathbf{y}$ zachodzi równość*

$$p(\mathbf{x}, \mathbf{y}) = p_{\mathbf{X}}(\mathbf{x}) p_{\mathbf{Y}}(\mathbf{y})\,.$$

Twierdzenie 2.13 (O niezależności funkcji zmiennych losowych)**.** *Niech $X_1, \ldots, X_n$ będą zmiennymi losowymi o wartościach $\mathcal{X}$, a $g_1, \ldots, g_n : \mathcal{X} \mapsto \mathcal{Y}$, $\varphi : \mathcal{X}^m \mapsto \mathcal{Y}$ i $\psi : \mathcal{X}^{n-m} \mapsto \mathcal{Y}$ funkcjami oraz $m < n$. Wówczas jeśli $X_1, \ldots, X_n$ są niezależne to:*

1. *zmienne losowe $g_1(X_1), \ldots, g_n(X_n)$ są niezależne;*
2. *zmienne losowe $\varphi(X_1, \ldots, X_m)$ i $\psi(X_{m+1}, \ldots, X_n)$ są niezależne.*

Jeśli $(\mathbf{X}, \mathbf{Y})$ jest zmienną losową o rozkładzie ciągłym z gęstością $p(\mathbf{x}, \mathbf{y})$ i gęstość brzegowa $p_{\mathbf{Y}}$ jest funkcją dodatnią, to dla każdego $\mathbf{y}$ rozkład warunkowy $\mathbf{X}$ pod warunkiem $\mathbf{Y} = \mathbf{y}$ także jest ciągły z gęstością

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p_{\mathbf{Y}}(\mathbf{y})}\,, \tag{2.7}$$

którą nazywamy gęstością warunkową $\mathbf{X}$ pod warunkiem $\mathbf{Y} = \mathbf{y}$.

Jeśli $(\mathbf{X}, \mathbf{Y})$ jest zmienną losową o rozkładzie dyskretnym z funkcją prawdopodobieństwa $p(\mathbf{x}, \mathbf{y})$ i brzegowa funkcja prawdopodobieństwa $p_{\mathbf{Y}}$ jest funkcją dodatnią, to dla każdego $\mathbf{y}$ rozkład warunkowy $\mathbf{X}$ pod warunkiem $\mathbf{Y} = \mathbf{y}$ także jest dyskretny z funkcją prawdopodobieństwa

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p_{\mathbf{Y}}(\mathbf{y})}\,, \tag{2.8}$$

którą nazywamy warunkową funkcją prawdopodobieństwa $\mathbf{X}$ pod warunkiem $\mathbf{Y} = \mathbf{y}$.

Definicja 2.36 (Warunkowej wartości oczekiwanej)**.** *Warunkową wartość oczekiwaną* $\mathbf{X}$ *pod warunkiem* $\mathbf{Y} = \mathbf{y}$ *nazywamy wielkość*

$$\mathbb{E}[\mathbf{X} \mid \mathbf{y}] = \int_{\mathbb{R}^n} \mathbf{x} p(\mathbf{x} \mid \mathbf{y}) \, \mathrm{d}^n \mathbf{x}$$

Definicja 2.37 (Warunkowej niezależności)**.** *Niech* $X_i : \Omega \mapsto \mathcal{X}$ *oraz* Y *będą zmiennymi losowymi. Mówimy, że zmienne* $X_1, \ldots, X_n$ *są warunkowo niezależne względem* Y*, jeśli dla dowolnych* $A_i \in \mathcal{F}_{\mathcal{X}}$ *i dowolnej wartości* y *zachodzi*

$$\begin{aligned} &P(X_1 \in A_1, \ldots, X_n \in A_n \mid Y = y) \\ &= P(X_1 \in A_1 \mid Y = y) \ldots P(X_n \in A_n \mid Y = y) \, . \end{aligned}$$

Dla rozkładów ciągłych lub dyskretnych z gęstością lub funkcją prawdopodobieństwa p warunkowa niezależność $\mathbf{X}, \mathbf{Y}$ względem $\mathbf{Z}$ jest równoważna

$$p(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = p_{\mathbf{X}}(\mathbf{x} \mid \mathbf{z}) p_{\mathbf{Y}}(\mathbf{y} \mid \mathbf{z}) \, .$$

2.8 Kowariancja i współczynnik korelacji

Definicja 2.38 (Kowariancji)**.** *Kowariancją zmiennych losowych rzeczywistych* X, Y *o nazywamy liczbę*

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - m_X)(Y - m_Y)] \, ,$$

gdzie wartości oczekiwane są liczone względem łącznego rozkładu prawdopodobieństwa $p(x, y)$*.* ***Współczynnikiem korelacji zmiennych*** X, Y *nazywamy liczbę*

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \, .$$

Twierdzenie 2.14 (Własności kowariancji)**.** *Niech* X, Y *będą zmiennymi losowymi rzeczywistymi, wówczas*

1. $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - m_X m_Y$*.*
2. *Jeśli* X *i* Y *są niezależne oraz istnieje* $\mathbb{E}(XY)$*, to* $\mathrm{Cov}(X, Y) = 0$*.*

Definicja 2.39 (Macierzy kowariancji)**.** *Macierzą kowariancji* $\mathbf{K}$ *n–wymiarowej zmiennej losowej* $\mathbf{X}$ *nazywamy macierz* $n \times n$ *zdefiniowaną jako*

$$\mathbf{K} = \mathbb{E}[(\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^T]$$

lub równoważnie

$$\mathbf{K}_{ij} = \begin{cases} \mathbb{V}[\mathbf{X}_i] \, , & i = j \\ \mathrm{Cov}(\mathbf{X}_i, \mathbf{X}_j) \, , & i \neq j \end{cases} \, .$$

Twierdzenie 2.15 (Własności macierzy kowariancji)**.** *Niech* $\mathbf{K}$ *będzie macierzą kowariancji zmiennej* $\mathbf{X}$*, wówczas*

1. *Macierz kowariancji jest symetryczna.*

2. *Macierz kowariancji jest nieujemnie określona.*

3. *Jeśli zmienne $\mathbf{X}_1, \ldots, \mathbf{X}_n$ są niezależne, to macierz kowariancji jest diagonalna.*

2.9 Wielowymiarowy rozkład normalny

Definicja 2.40 (Standardowego wielowymiarowego rozkładu normalnego)**.** *Zmienna losowa $\mathbf{X}$ ma standardowy n–wymiarowy rozkład normalny jeśli jej składowe są niezależne i dla każdego $i = 1, \ldots, n$ $\mathbf{X}_i \sim \mathcal{N}(0, 1)$. Jest to rozkład ciągły o gęstości*

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n}} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right).$$

Definicja 2.41 (Wielowymiarowego rozkładu normalnego)**.** *Zmienna losowa $\mathbf{X}$ ma n–wymiarowy rozkład normalny (z ang. Multivariate Normal Distribution, MVN), tzn. $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ jeśli istnieje k–wymiarowa zmienna losowa $\mathbf{Z}$ o standardowym rozkładzie normalnym dla pewnego $k \leq n$ oraz istnieje $\boldsymbol{\mu} \in \mathbb{R}^n$ i macierz $\mathbf{A} \in \mathbb{R}^{n \times k}$ takie, że $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$ oraz*

$$\mathbf{X} = \mathbf{A}\mathbf{Z} + \boldsymbol{\mu}.$$

Jeśli macierz $\boldsymbol{\Sigma}$ jest dodatnio określona, to rozkład $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ jest ciągły, a jego gęstość jest dana przez

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

*Macierz $\boldsymbol{\Sigma}^{-1}$ nazywa się **macierzą precyzji**.*

Poziomice gęstości niezdegenerowanego wielowymiarowego rozkładu normalnego są elipsoidami, których półosie są skierowane wzdłuż wektorów własnych macierzy $\boldsymbol{\Sigma}$ i mają długości proporcjonalne do pierwiastka z wartości własnych.

Twierdzenie 2.16 (Własności niezdegenerowanego rozkładu normalnego)**.** *Niech $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ dla dodatnio określonej macierzy $\boldsymbol{\Sigma}$, wówczas*

1. *Wszystkie rozkłady brzegowe i warunkowe $\mathbf{X}$ są rozkładami normalnymi.*

2. *Zmienne składowe $\mathbf{X}_1, \ldots, \mathbf{X}_n$ są niezależne wtedy i tylko wtedy, gdy $\boldsymbol{\Sigma}$ jest macierzą diagonalną.*

2.10 Rodzaje zbieżności zmiennych losowych

Niech $X_1, X_2, \ldots$ i X będą zmiennymi losowymi o wartościach w tej samej przestrzeni $\mathcal{X}$.

Definicja 2.42 (Zbieżności stochastycznej)**.** *Ciag (X_n) jest zbieżny do X według prawdopodobieństwa (lub stochastycznie) jeśli dla każdego $\epsilon > 0$*

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1\,.$$

Definicja 2.43 (Zbieżności z prawdopodobieństwem 1)**.** *Ciąg (X_n) jest zbieżny do X z prawdopodobieństwem 1 jeśli*

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1\,.$$

Definicja 2.44 (Zbieżności według dystrybuant)**.** *Załóżmy, że $\mathcal{X} = \mathbb{R}^k$ i niech $F_{\mathbf{X}^n}$ będzie dystrybuantą $\mathbf{X}^n$, $F_{\mathbf{X}}$ dystrybuantą $\mathbf{X}$. Ciąg $(\mathbf{X}^n)$ jest zbieżny do $\mathbf{X}$ według dystrybuant jeśli dla każdego $\mathbf{x} \in \mathbb{R}^k$ takiego, że $F_{\mathbf{X}}$ jest ciągła w $\mathbf{x}$ zachodzi*

$$\lim_{n\to\infty} F_{\mathbf{X}^n}(\mathbf{x}) = F_{\mathbf{X}}(\mathbf{x})\,.$$

Twierdzenie 2.17 (Zależności między rodzajami zbieżności)**.** *Niech (X_n) będzie ciągiem zmiennych losowych. Wówczas*

1. *Jeśli $X_n \to X$ z prawdopodobieństwem 1, to $X_n \to X$ według prawdopodobieństwa.*
2. *Jeśli $X_n \to X$ według prawdopodobieństwa, to istnieje podciąg X_{n_k} taki, że $X_{n_k} \to X$ z prawdopodobieństwem 1.*
3. *Jeśli $X_n \to X$ według prawdopodobieństwa, to $X_n \to X$ według dystrybuant.*

2.11 Wnioskowanie statystyczne

Niech zmienna losowa X określa model rozkładu pewnej cechy (cech) w ustalonym zbiorze instancji (tzw. **populacji generalnej**). Innymi słowy, przyjmujemy, że wartości cech zachowują się jakby zostały wybrane losowo zgodnie z rozkładem zmiennej X. Do podstawowych zagadnień wnioskowania statystycznego należą:

- oszacowanie wielkości charakteryzujących rozkład X (np. wartości średniej albo wariancji);
- weryfikacja hipotez dotyczących rozkładu X (tym nie będziemy się zajmować).

Definicja 2.45 (Modelu statystycznego)**.** *Modelem statystycznym nazywamy parę $(\mathcal{P}, \mathcal{X})$, gdzie $\mathcal{P}$ jest rodziną rozkładów prawdopodobieństwa na zbiorze $\mathcal{X}$. Zazwyczaj przyjmuje się*

$$\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$$

*dla pewnego zbioru parametrów Θ. Model statystyczny nazywamy **parametrycznym** jeśli $\Theta \subset \mathbb{R}^k$.*

Definicja 2.46 (Prostej próby losowej)**.** *Prostą próbą losową o liczności n nazywamy ciąg niezależnych zmiennych losowych $X_1, \ldots, X_n$ o wartościach w $\mathcal{X}$ i o tym samym rozkładzie $P_{\boldsymbol{\theta}} \in \mathcal{P}$ (z ang. independent and identically distributed, i.i.d).*

Definicja 2.47 (Realizacji prostej próby losowej)**.** *Ciąg wartości $x_1, \ldots, x_n \in \mathcal{X}$ takich, żeby*

$$X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n$$

dla pewnego ω nazywamy realizacją prostej próby losowej $X_1, \ldots, X_n$.

Definicja 2.48 (Statystyki)**.** *Statystyką nazywa się zmienną losową będącą funkcją prostej próby losowej $T(X_1, \ldots, X_n)$.*

Średnią w prostej próbie losowej nazywa się statystykę

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \,. \tag{2.9}$$

Wariancją w prostej próbie losowej nazywa się statystykę

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 \,. \tag{2.10}$$

2.12 Prawa wielkich liczb i CTG

Definicja 2.49. *Mówimy, że dla ciągu zmiennych losowych $X_1, X_2, \ldots$ zachodzi* ***słabe prawo wielkich liczb*** *jeżeli*

$$\frac{1}{n} \sum_{k=1}^{n} (X_k - \mathbb{E}[X_k]) \to 0$$

według prawdopodobieństwa. Jeżeli powyższy ciąg jest zbieżny z prawdopodobieństwem 1, to mówimy, że dla ciągu (X_n) zachodzi ***mocne prawo wielkich liczb****.*

Twierdzenie 2.18. *Słabe prawo wielkich liczb zachodzi, jeśli ciąg (X_n) spełnia* ***jeden*** *z podanych warunków:*

- $\frac{1}{n^2} \mathbb{V}\left[\sum_{k=1}^{n} X_k\right] \to 0$ *(prawo wielkich liczb Markowa).*
- *X_n są niezależne i ich wariancje są wspólnie ograniczone (prawo wielkich liczb Czebyszewa).*
- *X_n są niezależne i mają ten sam rozkład o skończonej wartości średniej (prawo wielkich liczb Chinczyna).*

Twierdzenie 2.19 (Prawo wielkich liczb Kołmogorowa)**.** *Jeśli X_n są niezależne, mają skończone wariancje oraz*

$$\sum_{n=1}^{\infty} \frac{\mathbb{V}[X_n]}{n^2} < +\infty\,,$$

to dla X_n zachodzi mocne prawo wielkich liczb (prawo wielkich liczb Kołmogorowa).

Twierdzenie 2.20 (Centralne Twierdzenie Graniczne)**.** *Niech X_n będzie ciągiem niezależnych zmiennych losowych o tym samym rozkładzie ze skończoną wartością średnią m i wariancją $\sigma^2 > 0$. Wtedy ciąg Z_n*

$$Z_n := \frac{\frac{1}{n}\sum_{k=1}^{n} X_k - m}{\frac{\sigma}{\sqrt{n}}}$$

jest zbieżny według dystrybuant do zmiennej losowej Z o standardowym rozkładzie normalnym.

2.13 Estymatory punktowe

Definicja 2.50 (Estymatora)**.** *Estymatorem nazywa się statystykę $\hat{\theta}(X_1, \ldots, X_n)$ służącą do oszacowania wartości parametru θ. Liczbę $\hat{\theta}(x_1, \ldots, x_n)$ dla konkretnej realizacji prostej próby losowej nazywa się wartością estymatora albo estymatą.*

Definicja 2.51 (Obciążenia estymatora)**.** *Liczbę*

$$\mathrm{B}(\hat{\theta}) = \mathbb{E}[\hat{\theta}(X_1, \ldots, X_n)] - \theta$$

nazywamy obciążeniem estymatora $\hat{\theta}$. Jeśli $\mathrm{B}(\hat{\theta}) = 0$ to estymator nazywamy ***nieobciążonym****.*

Definicja 2.52 (Błędu średniokwadratowego)**.** *Błędem średniokwadratowym (z ang. Mean Squared Error, MSE) estymatora $\hat{\theta}$ parametru θ nazywamy liczbę*

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]\,.$$

Twierdzenie 2.21. *Dla dowolnego estymatora $\hat{\theta}$ zachodzi*

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{V}[\hat{\theta}] + (\mathrm{B}(\hat{\theta}))^2\,.$$

Definicja 2.53 (Funkcji wiarygodności)**.** *Funkcją wiarygodności (z ang. likelihood function) dla modelu $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ nazywamy funkcję*

$$\mathcal{L} : \mathbb{R}^n \times \Theta \ni (\mathbf{x}, \boldsymbol{\theta}) \mapsto \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}) \in [0; +\infty)$$

wyznaczającą rozkład łączny obserwowanych danych jako funkcję parametru $\boldsymbol{\theta}$.

Niech $\mathbf{X}^1, \ldots, \mathbf{X}^n$ będzie prostą próbą losową. Jeśli $P_{\boldsymbol{\theta}}$ są rozkładami ciągłymi lub dyskretnymi o gęstościach lub funkcjach prawdopodobieństwa $p(\cdot; \boldsymbol{\theta})$, to

$$\mathcal{L}(\mathbf{x}^1, \ldots, \mathbf{x}^n; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}^i; \boldsymbol{\theta}) \,. \tag{2.11}$$

Dla wygody obliczeń często rozważa się tzw. zanegowaną logarytmiczną funkcję wiarygodności (z ang. *Negated Log-Likelihood function, NLL*), tzn.

$$L(\mathbf{x}; \boldsymbol{\theta}) = -\log \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}) \,. \tag{2.12}$$

Wówczas dla realizacji prostej próby losowej mamy

$$L(\mathbf{x}^1, \ldots, \mathbf{x}^n; \boldsymbol{\theta}) = -\sum_{i=1}^{n} \log p(\mathbf{x}^i; \boldsymbol{\theta}) \,. \tag{2.13}$$

Definicja 2.54 (Estymatora największej wiarygodności)**.** *Estymatorem największej wiarygodności (z ang. Maximum Likelihood Estimator, MLE) nazywamy funkcję $\hat{\boldsymbol{\theta}}$, która przy ustalonych wartościach obserwacji (realizacji prostej próby losowej) $\{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ maksymalizuje wartość funkcji wiarygodności lub, co równoważne, minimalizuje wartość zanegowanej logarytmicznej funkcji wiarygodności tj.*

$$\hat{\boldsymbol{\theta}}(\mathbf{x}^1, \ldots, \mathbf{x}^n) = \arg\min_{\boldsymbol{\theta} \in \Theta} \left[-\sum_{i=1}^{n} \log p(\mathbf{x}^i; \boldsymbol{\theta}) \right] .$$

Jeśli funkcja wiarygodności jest różniczkowalna względem $\boldsymbol{\theta}$ dla dowolnych $\mathbf{x}^i$, to MLE można czasem wyznaczyć analitycznie korzystając z warunku koniecznego optymalności, tzn. rozwiązując układ równań

$$\frac{\partial L(\mathbf{x}^1, \ldots, \mathbf{x}^n; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0 \,. \tag{2.14}$$

Jeśli MLE nie da się wyliczyć analitycznie, wyznacza się je przy użyciu algorytmów optymalizacji numerycznej. Estymatory MLE są asymptotycznie nieobciążone.

3 Elementy statystycznego uczenia maszynowego

Przechodzimy teraz do zagadnień uczenia maszynowego, w których wykorzystamy przedstawioną wcześniej teorię rachunku prawdopodobieństwa (w szczególności teorię zmiennych losowych) oraz wnioskowania statystycznego.

3.1 Podstawy wnioskowania bayesowskiego

Niech $\mathcal{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ będzie realizacją prostej próby losowej, czyli inaczej zbiorem obserwacji i.i.d. Zakładamy, że obserwacje te pochodzą z pewnego parametrycznego modelu statystycznego $\mathcal{P}$ z parametrami $\boldsymbol{\theta} \in \Theta$. Wcześniej $\boldsymbol{\theta}$

miał jedynie rangę parametru. We wnioskowaniu bayesowskim uznajemy, że parametry $\boldsymbol{\theta}$ są również zmiennymi losowymi, a model statystyczny modeluje warunkowy rozkład prawdopodobieństwa obserwacji pod warunkiem parametru. Mamy zatem rodzinę gęstości prawdopodobieństwa $p(\mathbf{x} \mid \boldsymbol{\theta})$ i chcemy wnioskować o parametrze $\boldsymbol{\theta}$ na podstawie obserwacji $\mathcal{X}$. Jeśli znamy rozkład a priori (inaczej **prior**) parametru $\boldsymbol{\theta}$ opisany przez $p(\boldsymbol{\theta})$, to z twierdzenia Bayesa rozkład a posteriori (**posterior**) jest dany przez

$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{X})} = \frac{p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}' \in \Theta} p(\mathcal{X} \mid \boldsymbol{\theta}')p(\boldsymbol{\theta}')} \tag{3.1}$$

Nie jesteśmy tutaj zbyt formalni z notacją, gdyż używamy p bardziej jakby był to rozkład prawdopodobieństwa (miara probabilistyczna), a w rzeczywistości jest to gęstość lub funkcja prawdopodobieństwa, więc powinniśmy używać indeksów dolnych określających zmienną losową, której rozkład jest opisany danym p, aby rozróżnić człony od siebie. Zapis taki jest jednak niezwykle wygodny i dość czytelny. Należy jedynie pamiętać, iż nazwa argumentu funkcji p określa teraz z jakiego wzoru powinniśmy skorzystać aby obliczyć jej wartość. Człon w mianowniku postaci

$$Z = \sum_{\boldsymbol{\theta}' \in \Theta} p(\mathcal{X} \mid \boldsymbol{\theta}')p(\boldsymbol{\theta}') \approxeq \int_{\Theta} p(\mathcal{X} \mid \boldsymbol{\theta}')p(\boldsymbol{\theta}') \,\mathrm{d}^n\boldsymbol{\theta}' \tag{3.2}$$

jest tzw. **czynnikiem normalizacyjnym** (czyli po prostu liczbą, często oznaczaną przez Z), który zapewnia, iż $p(\boldsymbol{\theta} \mid \mathcal{X})$ sumuje / całkuje się do 1.

Ponieważ założyliśmy, iż obserwacje ze zbioru $\mathcal{X}$ są warunkowo niezależne względem parametru $\boldsymbol{\theta}$ oraz pochodzą z tego samego rozkładu opisanego przez $p(\mathbf{x} \mid \boldsymbol{\theta})$, więc człon $p(\mathcal{X} \mid \boldsymbol{\theta})$ zwany **wiarygodnością** możemy zapisać jako

$$p(\mathcal{X} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}^i \mid \boldsymbol{\theta})\,. \tag{3.3}$$

Całe wnioskowanie bayesowskie opiera się na wyznaczeniu rozkładu a posteriori dla danego zbioru obserwacji $\mathcal{X}$, który wyraża naszą wiedzę o estymowanym parametrze $\boldsymbol{\theta}$. Na podstawie tego rozkładu możemy wyznaczyć estymatę punktową MAP maksymalizującą gęstość prawdopodobieństwa a posteriori,

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}(\mathcal{X}) = \arg\max_{\boldsymbol{\theta} \in \Theta} p(\boldsymbol{\theta} \mid \mathcal{X})\,, \tag{3.4}$$

jak również niepewność związaną z wyznaczeniem tej estymaty np. poprzez wyznaczenie przedziału wiarygodności (nie należy mylić z przedziałem ufności). Możemy również skonstruować rozkład predykcyjny (z ang. *posterior predictive distribution*) określający prawdopodobieństwo uzyskania nowej obserwacji $\mathbf{t}$

$$p(\mathbf{t} \mid \mathcal{X}) = \int_{\Theta} p(\mathbf{t} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{X}) \,\mathrm{d}^n\boldsymbol{\theta}\ . \tag{3.5}$$

Znając rozkład a posteriori estymowanego parametru $\boldsymbol{\theta}$ możemy nie tylko wyznaczyć estymaty punktowe, wartości oczekiwane i przedziały wiarygodności, ale również znaleźć estymator Bayesa (z ang. *Bayes estimator*), który minimalizuje wartość oczekiwaną pewnej funkcji straty (z ang. *loss function*) $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ po wszystkich estymatorach $\hat{\boldsymbol{\theta}}$

$$\hat{\boldsymbol{\theta}}_{\text{Bayes}}(\mathcal{X}) = \arg\min_{\hat{\boldsymbol{\theta}} \in \Theta} \int_{\Theta} L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) p(\boldsymbol{\theta} \mid \mathcal{X}) \, \mathrm{d}^n \boldsymbol{\theta} \, . \tag{3.6}$$

Całkę w powyższym wzorze nazywa się również funkcją ryzyka (z ang. *risk function*) $R(\hat{\boldsymbol{\theta}})$, która określa oczekiwaną stratę spowodowaną wykorzystaniem danego estymatora parametru.

Są dwa zasadnicze problemy we wnioskowaniu bayesowskim: pierwszym jest potrzeba znania rozkładu a priori estymowanego parametru, drugim – problem z obliczeniem czynnika normalizującego, który może być skomplikowaną całką lub sumą po wykładniczo-wielu elementach. Oba te problemy można czasami rozwiązać wprowadzając tzw. **prior sprzężony do wiarygodności**, tzn. zakładamy taki rozkład a priori, aby dla danej wiarygodności rozkład a posteriori miał znaną postać (np. rozkładu normalnego, rozkładu beta), wówczas nie musimy obliczać czynnika normalizującego, gdyż jest on po prostu znany.

3.2 Modele gaussowskie i liniowe modele gaussowskie

Zajmiemy się teraz wnioskowaniem bayesowskim w modelach, w których potrafimy analitycznie znaleźć postać rozkład a posteriori. Jak już wspomnieliśmy (Tw. 2.16) gdy zmienna losowa $\mathbf{X}$ ma wielowymiarowy rozkład normalny z wartością oczekiwaną $\boldsymbol{\mu}$ i dodatnio określoną macierzą kowariancji $\boldsymbol{\Sigma}$ to wszystkie rozkłady warunkowe i brzegowe są rozkładami normalnymi. Poniżej wyznaczymy parametry tych rozkładów.

Twierdzenie 3.1. *Niech zmienne losowe $\mathbf{x} \in \mathbb{R}^{n-k}$ i $\mathbf{y} \in \mathbb{R}^k$ mają łącznie wielowymiarowy rozkład normalny*

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{y}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}} \\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}} \end{bmatrix} \right) .$$

Wówczas

1. *Zmienne losowe $\mathbf{x}$ i $\mathbf{y}$ mają odpowiednio rozkłady*

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{xx}}) \, , \quad \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{yy}}) \, .$$

2. *Rozkład warunkowy $\mathbf{x} \mid \mathbf{y}$ jest rozkładem normalnym $\mathbf{x} \mid \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}})$ o parametrach*

$$\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\mu}_{\mathbf{x}} + \boldsymbol{\Sigma}_{\mathbf{xy}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} (\mathbf{y} - \boldsymbol{\mu}_{\mathbf{y}}) \, , \quad \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} = \boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{xy}} \boldsymbol{\Sigma}_{\mathbf{yy}}^{-1} \boldsymbol{\Sigma}_{\mathbf{yx}} \, .$$

Powyższe własności normalnych rozkładów łącznych pozwalają jawnie wnioskować w tzw. liniowych modelach gaussowskich (z ang. *Linear Gaussian Models*). Załóżmy, iż nasze obserwacje są modelowane zmienną losową $\mathbf{y}$ o rozkładzie normalnym z estymowanym parametrem $\mathbf{x}$ i znanymi parametrami $\mathbf{A}, \mathbf{b}, \mathbf{\Sigma_y}$ tak, że wiarygodność ma postać

$$\mathbf{y} \mid \mathbf{x} \sim \mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{\Sigma_y}) \,, \tag{3.7}$$

gdzie $\mathbf{A}$ jest w ogólności macierzą prostokątną. Jeśli jako prior na $\mathbf{x}$ przyjmiemy również rozkład normalny

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{x}, \mathbf{\Sigma_x}) \tag{3.8}$$

to rozkład a posteriori jest również rozkładem normalnym. W szczególności załóżmy, że mamy zbiór obserwacji i.i.d. $\mathcal{X} = \{\mathbf{y}^1, \ldots, \mathbf{y}^n\}$. Wówczas wiarygodność ma postać

$$p(\mathcal{X} \mid \mathbf{x}) \propto \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(\mathbf{y}^i - (\mathbf{A}\mathbf{x} + \mathbf{b}))^T \mathbf{\Sigma}_\mathbf{y}^{-1} (\mathbf{y}^i - (\mathbf{A}\mathbf{x} + \mathbf{b}))\right) , \tag{3.9}$$

a rozkład a posteriori

$$p(\mathbf{x} \mid \mathcal{X}) \propto p(\mathcal{X} \mid \mathbf{x}) \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^T \mathbf{\Sigma}_\mathbf{x}^{-1} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})\right) . \tag{3.10}$$

Rozpisując wszystkie czynniki i pomijając czynnik stałe otrzymujemy

$$\begin{aligned} \log p(\mathbf{x} \mid \mathcal{X}) \propto & -\frac{1}{2}\mathbf{x}^T \left(\mathbf{\Sigma}_\mathbf{x}^{-1} + n\mathbf{A}^T \mathbf{\Sigma}_\mathbf{y}^{-1} \mathbf{A}\right) \mathbf{x} \\ & + \mathbf{x}^T \left(\mathbf{\Sigma}_\mathbf{x}^{-1} \boldsymbol{\mu}_\mathbf{x} + \mathbf{A}^T \mathbf{\Sigma}_\mathbf{y}^{-1} \left[\sum_{i=1}^{n} \mathbf{y}^i - n\mathbf{b}\right]\right) \end{aligned} \tag{3.11}$$

skąd widzimy, iż rozkład $\mathbf{x} \mid \mathcal{X}$ jest rozkładem normalnym o parametrach

$$\boxed{\begin{aligned} \boldsymbol{\mu}_{\mathbf{x}|\mathcal{X}} &= \mathbf{\Sigma}_{\mathbf{x}|\mathcal{X}} \left(\mathbf{\Sigma}_\mathbf{x}^{-1} \boldsymbol{\mu}_\mathbf{x} + \mathbf{A}^T \mathbf{\Sigma}_\mathbf{y}^{-1} \left[\sum_{i=1}^{n} \mathbf{y}^i - n\mathbf{b}\right]\right) , \\ \mathbf{\Sigma}_{\mathbf{x}|\mathcal{X}} &= \left(\mathbf{\Sigma}_\mathbf{x}^{-1} + n\mathbf{A}^T \mathbf{\Sigma}_\mathbf{y}^{-1} \mathbf{A}\right)^{-1} . \end{aligned}} \tag{3.12}$$

3.3 Regresja liniowa

Załóżmy, iż modelujemy obserwacje postaci $(y, \mathbf{x})$, gdzie y to skalar zwany **zmienną objaśnianą**, którego wartość obserwujemy, a $\mathbf{x}$ to wektor zmiennych objaśniających, co do którego zakładamy, iż dla danego pomiaru y jest on znany dokładnie. Dodatkowo załóżmy liniowy model $\hat{y}(\mathbf{x}; \mathbf{w})$ postaci

$$\hat{y}(\mathbf{x}; \mathbf{w}) = \mathbf{x}^T \mathbf{w} \,, \tag{3.13}$$

w którym mamy dodatkowo **błąd losowy** $\epsilon \sim \mathcal{N}(0, \sigma^2)$ z nieznanym σ. Możemy zatem napisać model statystyczny postaci

$$y(\mathbf{x}) \mid \mathbf{w}, \sigma \sim \mathcal{N}(\hat{y}(\mathbf{x}; \mathbf{w}), \sigma^2)\,, \tag{3.14}$$

gdzie $\mathbf{w}$, σ to estymowane parametry. Powiedzmy, iż mamy zbiór obserwacji i.i.d. $\mathcal{X} = \{y_1(\mathbf{x}^1), \ldots, y_n(\mathbf{x}^n)\}$. Wiarygodność ma zatem postać

$$\boxed{\mathcal{L}(\mathcal{X}; \mathbf{w}, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\left[y_i - \hat{y}(\mathbf{x}^i; \mathbf{w})\right]^2\right)\,.} \tag{3.15}$$

W przypadku regresji liniowej zamiast pełnego wnioskowania bayesowskiego często stosuje się prostsze podejście polegające na ograniczeniu się do znalezienia estymaty punktowej MLE. Zanegowana logarytmiczna funkcja wiarygodności, którą będziemy również nazywać **funkcją kosztu** ma postać (pomijamy człony stałe, gdyż nie są one istotne przy dalszej minimalizacji)

$$\boxed{L(\mathcal{X}; \mathbf{w}, \sigma) = n \log \sigma + \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[y_i - \hat{y}(\mathbf{x}^i; \mathbf{w})\right]^2 + \text{const.}} \tag{3.16}$$

Minimalizując funkcję L względem $\mathbf{w}$ i σ otrzymamy estymaty MLE tych parametrów. Dla ustalonego stałego σ otrzymana funkcja L ma postać formy kwadratowej i otrzymany przy takim uproszczeniu problem optymalizacyjny nazywamy metodą najmniejszych kwadratów (z ang. *Ordinary Least Squares, OLS*). W przypadku modelu liniowego estymatory można znaleźć analitycznie rozwiązując układ równań

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}_j} &= \frac{1}{2\sigma^2} \frac{\partial}{\partial \mathbf{w}_j} \sum_{i=1}^{n} \left[y_i - \mathbf{w}^T \mathbf{x}^i\right]^2 = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}^i) \mathbf{x}_j^i = 0\,, \\ \frac{\partial L}{\partial \sigma} &= \frac{n}{\sigma} - \frac{1}{\sigma^3} \sum_{i=1}^{n} \left[y_i - \hat{y}(\mathbf{x}^i; \mathbf{w})\right]^2 = 0\,. \end{aligned} \tag{3.17}$$

Z powyższego

$$\begin{aligned} &\sum_{i=1}^{n} y_i \mathbf{x}_j^i - \mathbf{w}_{\text{MLE}}^T \sum_{i=1}^{n} \mathbf{x}^i \mathbf{x}_j^i = 0\,, \\ &\sigma_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \hat{y}(\mathbf{x}^i; \mathbf{w}_{\text{MLE}})\right]^2\,. \end{aligned} \tag{3.18}$$

Wprowadzając wektor $\mathbf{y}_i := y_i$ oraz macierz $\mathbf{X}_{ij} := \mathbf{x}_j^i$ możemy zapisać pierwsze równanie jako

$$-\mathbf{y}^T \mathbf{X} + \mathbf{w}_{\text{MLE}}^T \mathbf{X}^T \mathbf{X} = 0\,, \tag{3.19}$$

skąd

$$\boxed{\begin{aligned}\mathbf{w}_{\mathrm{MLE}} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^+\mathbf{y}\,,\\ \sigma^2_{\mathrm{MLE}} &= \frac{1}{n}\left(\mathbf{y}-\mathbf{X}\mathbf{w}_{\mathrm{MLE}}\right)^T\left(\mathbf{y}-\mathbf{X}\mathbf{w}_{\mathrm{MLE}}\right)\,,\end{aligned}} \tag{3.20}$$

gdzie $\mathbf{X}^+$ oznacza **pseudoodwrotność Moore'a–Penrose'a**, którą można efektywnie obliczyć korzystając z rozkładu SVD macierzy $\mathbf{X}$.

3.4 Regularyzacja

Regularyzacją nazywamy proces polegający na wprowadzeniu ad hoc do zagadnienia optymalizacji dodatkowych członów tak, aby rozwiązanie było regularne (prostsze, nieosobliwe, jednoznaczne). W przypadku funkcji kosztu L najczęściej dodajemy człon penalizujący rozwiązania o dużej normie estymowanego parametru tj. człon postaci $\gamma\|\mathbf{w}\|$ dla pewnej normy $\|\cdot\|$ i hiperparametru γ. W kontekście bayesowskim regularyzację można również rozumieć jako pewną „niechęć" (tłumienie, zachowawczość) modelu do zmiany rozkładu a priori estymowanego parametru.

W przypadku regresji liniowej jeśli zamiast poszukiwania estymaty MLE będziemy poszukiwać estymaty MAP (z ang. *Maximum a Posteriori estimate*) z rozkładem a priori na parametr $\mathbf{w}$ danym przez $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{1})$, to logarytm gęstości rozkładu a posteriori (który również będziemy nazywać zregularyzowaną funkcją kosztu) ma postać (tutaj zakładamy, że σ jest znaną stałą)

$$\boxed{L(\mathcal{X};\mathbf{w}) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - \hat{y}(\mathbf{x}^i;\mathbf{w})\right]^2 + \frac{1}{2\tau^2}\mathbf{w}^T\mathbf{w} + \text{const.}} \tag{3.21}$$

skąd możemy bez problemy wyznaczyć estymatę punktową MAP parametru $\mathbf{w}$

$$\mathbf{w}_{\mathrm{MAP}} = \left(\gamma\mathbf{1} + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\,, \tag{3.22}$$

gdzie $\gamma = \frac{\sigma^2}{\tau^2}$ nazywamy **siłą regularyzacji**. Zagadnienie minimalizacji funkcji kosztu będącej formą kwadratową z dodanym członem regularyzującym w postaci sumy kwadratów współrzędnych wektora $\mathbf{w}$ (normy L2 wektora) nazywamy **regresją grzbietową** (z ang. *ridge regression*), natomiast taką postać członu regularyzującego – regularyzacją L2. Zauważmy, że im większa jest wartość γ (mniejsza niepewność związana z rozkładem a priori) tym drugi człon w nawiasie staje się mniej istotny.

Innym przykładem regularyzacji jest tzw. regularyzacja L1, która polega na dodaniu do funkcji kosztu członu postaci $\gamma\sum_{j=1}^{d}|\mathbf{w}_j|$ tj. normy L1 wektora wag. Zagadnie optymalizacji formy kwadratowej z członem regularyzującym L1 nazywamy **regresją LASSO**. W takim przypadku nie da się prosto analitycznie znaleźć estymaty punktowej MAP i trzeba używać algorytmów optymalizacji numerycznej. W ogólności można połączyć regularyzacje L1 i L2 tj. rozważać zregularyzowaną funkcję kosztu postaci (tutaj parametr σ nie występuje, tj. ta

funkcja kosztu nie ma bezpośredniej interpretacji probabilistycznej jako funkcja wiarygodności)

$$\boxed{L(\mathcal{X};\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}\left[y_i - \hat{y}(\mathbf{x}^i;\mathbf{w})\right]^2 + \frac{\gamma_1}{2}\|\mathbf{w}\|_1 + \frac{\gamma_2}{2}\|\mathbf{w}\|_2 .} \tag{3.23}$$

Zagadnienie minimalizacji takiej funkcji kosztu nazywamy ElasticNet i tak jak w przypadku LASSO musimy korzystać z algorytmów optymalizacji numerycznej. Często wykorzystuje się tutaj algorytmy bezgradientowe np. coordinate descent.

3.5 Robust regression

Jednym z problemów wynikających z modelowania zależności $y(\mathbf{x})$ przez rozkład normalny jest duża czułość takiego modelu na wartości odstające. Wynika to z tego, iż wynikająca z takiego modelu funkcja kosztu jest proporcjonalna do kwadratu odległości punktu od prostej regresji. Metody robust to zbiór modeli odpornych na wartości odstające, które to modele przypisują im mniejszą wagę niż standardowa regresja liniowa. Jednym z modeli robust jest tzw. **regresja LAD** (z ang. *Least Absolute Deviation regression*), która (w ujęciu bayesowskim) modeluje zależność przez rozkład Laplace'a, który posiada tzw. ciężkie ogony (z ang. *heavy tails*) i dzięki temu przypisuje większe prawdopodobieństwo wartościom odstającym, co sprawia, że model jest na nie mniej czuły. Rozkład Laplace'a przekłada się na zmianę w funkcji kosztu kwadratów odległości na wartość bezwzględną odległości.

Innym przykładem robust regression jest **regresja Hubera**, w której heurystycznie wprowadzamy funkcję kosztu Hubera (z ang. *Huber loss*) postaci

$$L(\mathcal{X};\mathbf{w}) = \sum_{i=1}^{n}\ell_{\mathrm{H}}\left[y_i, \hat{y}(\mathbf{x}^i;\mathbf{w})\right] , \tag{3.24}$$

gdzie

$$\ell_{\mathrm{H}}\left[y, \hat{y}(\mathbf{x};\mathbf{w})\right] := \begin{cases} \frac{1}{2}\left[y - \hat{y}(\mathbf{x};\mathbf{w})\right]^2 , & |y - \hat{y}(\mathbf{x};\mathbf{w})| \le \epsilon \\ -\frac{1}{2}\epsilon^2 + \epsilon|y - \hat{y}(\mathbf{x};\mathbf{w})| , & |y - \hat{y}(\mathbf{x};\mathbf{w})| \ge \epsilon \end{cases} , \tag{3.25}$$

przy czym ϵ jest pewnym hiperparametrem. Widzimy, że koszt Hubera został tak wybrany aby być ciągłym i różniczkowalnym połączeniem błędów L1 i L2. Oczywiście możemy dodawać tutaj również człony regularyzujące (najczęściej L2).

4 Uczenie głębokie i sieci neuronowe