Spis treści

1 Rachunek prawdopodobieństwa

1.1 Definicja przestrzeni probabilistycznej

Rozkładem prawdopodobieństwa P w pewnym zbiorze zdarzeń elementarnych $\Omega \neq \emptyset$ nazywamy odwzorowanie

$$P : \Sigma \mapsto [0; 1] \,,$$

gdzie Σ jest rodziną podzbiorów Ω (inaczej rodziną zdarzeń) taką, że

$$\Omega \in \Sigma \,, \quad A \in \Sigma \implies A' \in \Sigma \,, \quad \forall A_1, A_2, \ldots \in \Sigma : \bigcup_i A_i \in \Sigma \,,$$

które spełnia: $P(\Omega) = 1$ oraz dla dowolnych parami rozłącznych zdarzeń $A_1, A_2, \ldots \in \Sigma$ zachodzi

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i) \,.$$

Trójkę (Ω, Σ, P) nazywamy przestrzenią probabilistyczną. Z powyższej definicji wynikają znane własności prawdopodobieństwa tj. $P(A') = 1 - P(A)$ oraz $P(A \cup B) = P(A) + P(B) - P(A, B)$.

1.2 Prawdopodobieństwo warunkowe

Definiujemy również prawdopodobieństwo warunkowe zdarzenia A pod warunkiem zdarzenia B o dodatnim prawdopodobieństwie

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

Na podstawie powyższej definicji definiujemy niezależność zdarzeń A, B jako własność $P(A, B) = P(A)P(B)$, co dla zdarzenia B o dodatnim prawdopodobieństwie jest równoważne z $P(A \mid B) = P(A)$. Ponadto jeśli zdarzenia $A_1, A_2, \ldots \in \Sigma$ są parami rozłączne i zachodzi $\bigcup_i A_i = \Omega$ to dla dowolnego zdarzenia $B \in \Sigma$ możemy zapisać

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

Z definicji prawdopodobieństwa warunkowego trywialnie udowodnić twierdzenie Bayesa

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

1.3 Zmienne losowe

W uczeniu maszynowym będą interesować nas zmienne o wartościach w $\mathbb{R}^n$. Zmienne takie nazywamy zmiennymi losowymi wielowymiarowymi i definiujemy jako odwzorowania

$$X : \Omega \mapsto \mathbb{R}^n$$

takie, że dla każdego $A \subseteq \mathbb{R}^n$ zbiór $\{\omega \in \Omega \mid X(\omega) \in A\}$ należy do rodziny zdarzeń Σ. Przy takiej definicji prawdopodobieństwo, iż zmienna X ma wartość należącą do pewnego przedziału A wynosi

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

Dowolny rozkład prawdopodobieństwa zmiennej losowej n–wymiarowej $X = (X_1, X_2, \ldots, X_n)$ jest wyznaczony jednoznacznie przez zadanie funkcji $F(\mathbf{x}) : \mathbb{R}^n \mapsto [0; 1]$ zwanej dystrybuantą zdefiniowanej jako

$$F(\mathbf{x}) = F(x_1, \ldots, x_n) := P(X_1 \leq x_1, \ldots, X_n \leq x_n) \,.$$

Zasadniczo będą nas interesować jednak dwa przypadki rozkładów prawdopodobieństwa zmiennych losowych: rozkłady dyskretne i rozkłady ciągłe. W przypadku rozkładu dyskretnego istnieje pewien przeliczalny zbiór $S \subset \mathbb{R}^n$ taki, że $P(X \in S) = 1$. Rozkład ten jest zadany jednoznacznie przez podanie $|S|$ liczb $p_i > 0$ określających prawdopodobieństwa $p_i = P(X = \mathbf{x}_i)$ dla wszystkich $\mathbf{x}_i \in S$. W przypadku rozkładu ciągłego istnieje z kolei funkcja $p(\mathbf{x}) : \mathbb{R}^n \mapsto [0; \infty)$ taka, że

$$P(X_1 \in [a_1; b_1], \ldots, X_n \in [a_n; b_n]) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} p(\mathbf{x}) \, \mathrm{d}^n\mathbf{x} \,.$$

Funkcje $p(\mathbf{x})$ nazywamy gęstością prawdopodobieństwa. W obu przypadkach musi być spełniony warunek unormowania postaci odpowiednio

$$\sum_i p_i = 1 \,, \quad \int_{\mathbb{R}^n} p(\mathbf{x}) \, \mathrm{d}^n\mathbf{x} = 1 \,.$$

Będziemy często wykorzystywać wartość oczekiwaną pewnej funkcji $f(\mathbf{x})$ zmiennej losowej X zdefiniowaną odpowiednio dla rozkładu p – dyskretnego lub ciągłego jako

$$\mathbb{E}[f(\mathbf{x})] := \sum_{\mathbf{x}_i \in S} f(\mathbf{x}_i) p_i \cong \int_{\mathbb{R}^n} f(\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}^n\mathbf{x} \,.$$

Zauważmy przy tym, iż funkcja $f(\mathbf{x})$ może być zupełnie dowolna, np. dla funkcji charakterystycznej (indykatorowej) zbioru $A \subset \mathbb{R}^n$ $f(\mathbf{x}) = \mathcal{I}_A$ mamy $\mathbb{E}[\mathcal{I}_A(\mathbf{x})] = P(X \in A)$ lub dla iloczynu funkcji Heaviside'a $f(\mathbf{x}) = \theta(t_1 - x_1) \cdots \theta(t_n - x_n)$ mamy $\mathbb{E}[f(\mathbf{x})] = F(t_1, \ldots, t_n)$.

1.4 Rozkłady brzegowe

Niech $X = (X_1, \ldots, X_n)$ będzie n–wymiarową zmienną losową o dystrybuancie $F(\mathbf{x})$. Rozkład brzegowy względem k zmiennych $X_{\sigma(1)}, \ldots, X_{\sigma(k)}$ definiujemy jako rozkład wyznaczony przez dystrybuantę

$$F_{X_{\sigma(1)},\ldots,X_{\sigma(k)}}(x_{\sigma(1)}, \ldots, x_{\sigma(k)}) := \lim_{x_{\sigma(k+1)}\to\infty,\ldots,x_{\sigma(n)}\to\infty} F(x_1, \ldots, x_n)\,.$$

1.5 Zmienne losowe niezależne

Niech $X = (X_1, \ldots, X_k)$ będzie n–wymiarową zmienną losową o rozkładzie wyznaczonym przez dystrybuantę $F(\mathbf{x})$. Powiemy, iż zmienne losowe $n_1, \ldots, n_k$ – wymiarowych $(n_1 + \ldots + n_k = n)$ $X_1, \ldots, X_k$ są niezależne iff dla dowolnych $\mathbf{x}_1 \in \mathbb{R}^{n_1}, \ldots, \mathbf{x}_k \in \mathbb{R}^{n_k}$ zachodzi

$$F(\mathbf{x}_1, \ldots, \mathbf{x}_k) = F_{X_1}(\mathbf{x}_1) \cdot \ldots \cdot F_{X_k}(\mathbf{x}_k)\,.$$

1.6 Funkcja tworząca momenty

Funkcją tworzącą momenty (z ang. *moment generating function*) nazywamy funkcję określoną wzorem

$$M_X(t) := \mathbb{E}\left[\mathrm{e}^{tX}\right]$$

dla zmiennej losowej rzeczywistej $X : \Sigma \mapsto \mathbb{R}$. Powyższa wielkość jest określona zawsze tylko dla $t = 0$. Dla innych t, $M_X(t)$ może w ogólności nie istnieć. MGF używamy, aby łatwiej obliczać momenty różnych rzędów. Istotnie jeśli $M_X(t)$ istnieje w pewnym otoczeniu $t = 0$ to

$$\mathbb{E}[X^k] = \frac{\mathrm{d}^k M_X}{\mathrm{d}t^k}\Big|_{t=0}\,.$$

Jednocześnie łatwo pokazać, że zachodzi $M_X(at) = M_{aX}(t)$ oraz $M_{X+b}(t) = M_X(t)\mathrm{e}^{tb}$.

1.7 Funkcja charakterystyczna

Funkcję charakterystyczną rozkładu o gęstości $p(x)$ definiujemy jako

$$\varphi_X(x) = \mathbb{E}\left[\mathrm{e}^{\mathrm{i}tx}\right] = \int_{-\infty}^{+\infty} p(x)\mathrm{e}^{\mathrm{i}tx}\,\mathrm{d}x\ .$$

Widzimy, iż jest to zatem transformata Fouriera funkcji gęstości. Funkcja charakterystyczna koduje pełną informację o rozkładzie i możemy wyciągną z niej funkcję gęstości przez zastosowanie odwrotnej transformacji Fouriera.

1.8 Rozkłady warunkowe

W ogólnym przypadku zmiennej losowej n – wymiarowej $Z = (Z_1, \ldots, Z_n)$ o ciągłym rozkładzie $p(\mathbf{z})$ jeśli wydzielimy zmienne k i $n-k$ – wymiarowe $X = (Z_{\sigma(1)}, \ldots, Z_{\sigma(k)})$, $Y = (Z_{\sigma(k+1)}, \ldots, Z_{\sigma(n)})$ to rozkład warunkowy zmiennej $X \mid Y$ definiujemy jako rozkład zadany przez gęstość prawdopodobieństwa

$$p(\mathbf{x} \mid \mathbf{y}) := \frac{p(\mathbf{z})}{p_Y(\mathbf{y})} = \frac{p(\mathbf{x}, \mathbf{y})}{p_Y(\mathbf{y})} \,.$$

1.9 Transformacja zmiennych wielowymiarowych

Niech $X = (X_1, \ldots, X_n)$ będzie zmienną losową wielowymiarową o rozkładzie ciągłym o gęstości $p_X(\mathbf{x})$. Rozważmy bijekcję $(X_1, \ldots, X_n) \mapsto (Y_1, \ldots, Y_n)$. Chcemy znaleźć wyrażenie na gęstość $p_Y(\mathbf{y})$ w nowych zmiennych. Ponieważ infinitezymalne prawdopodobieństwo jest niezmiennicze względem zmiany współrzędnych więc zachodzi

$$p_X(x_1, \ldots, x_n)\, \mathrm{d}x_1 \ldots \mathrm{d}x_n = p_Y(y_1, \ldots, y_n)\, \mathrm{d}y_1 \ldots \mathrm{d}y_n \ ,$$

skąd

$$p_Y(y_1, \ldots, y_n) = \left| \frac{\partial(x_1, \ldots, x_n)}{\partial(y_1, \ldots, y_n)} \right| p_X(x_1(\mathbf{y}), \ldots, x_n(\mathbf{y})) \,.$$

1.10 Macierz kowariancji

Macierz kowariancji funkcji $f(\mathbf{x})$ zmiennej losowej X definiujemy jako

$$\mathbf{\Sigma}[f(\mathbf{x})] := \mathbb{E}\left[(f(\mathbf{x}) - \boldsymbol{\mu}_f)(f(\mathbf{x}) - \boldsymbol{\mu}_f)^\top \right] \ ,$$

gdzie $\boldsymbol{\mu}_f = \mathbb{E}[f(\mathbf{x})]$. Elementy diagonalne Σ_{ii} tej macierzy nazywamy wariancjami zmiennych X_i, natomiast elementy pozadiagonalne Σ_{ij} nazywamy kowariancjami zmiennych X_i i X_j. Oczywiście $\mathbf{\Sigma}$ jest macierzą symetryczną. Nadto jeśli f jest funkcją identycznościową tj. $f(\mathbf{x}) = \mathbf{x}$ to $\mathbf{\Sigma}$ jest macierzą nieujemnie określoną, gdyż dla dowolnego $\mathbf{v} \in \mathbb{R}^n$ mamy

$$\mathbf{v}^\top \mathbf{\Sigma} \mathbf{v} = \mathbb{E}[\mathbf{v}^\top (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{v}] = \mathbb{E}[z^2] \geq 0 \,,$$

gdzie $z = \mathbf{v}^\top (\mathbf{x} - \boldsymbol{\mu}) \in \mathbb{R}$. Jeśli $X_1, \ldots, X_n$ są niezależne i f jest funkcją identycznościową to $\mathbf{\Sigma}$ jest macierzą diagonalną.

1.11 Wielowymiarowy rozkład normalny

Jeśli zmienna wielowymiarowa $X = (X_1, \ldots, X_n)$ ma wielowymiarowy rozkład normalny (z ang. *Multivariate Normal distribution – MVN*) z wartością oczekiwaną $\boldsymbol{\mu}$ i macierzą kowariancji $\boldsymbol{\Sigma}$, co oznaczamy jako $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, to gęstość prawdopodobieństwa jest dana

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

Macierz $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ nazywamy macierzą precyzji. Jeśli $\mathbf{v}_i$ są unormowanymi wektorami własnymi macierzy $\boldsymbol{\Sigma}$, a λ_i odpowiadającymi im wartościami własnymi i zakładając, iż widmo $\{\lambda_i\}$ jest niezdegenerowane mamy z twierdzenia spektralnego

$$\boldsymbol{\Lambda} = \sum_{i=1}^{n} \frac{1}{\lambda_i} \mathbf{v}_i \mathbf{v}_i^\top$$

oraz wiemy, iż wektory $\{\mathbf{v}_i\}$ tworzą bazę ortonormalną przestrzeni $\mathbb{R}^n$. Z powyższego możemy zatem wyrazić wektor $\mathbf{x}-\boldsymbol{\mu}$ jako kombinację liniową wektorów $\{\mathbf{v}_i\}$ tj.

$$\mathbf{x} - \boldsymbol{\mu} = \sum_{i=1}^{n} t_i \mathbf{v}_i \,,$$

co pozwala zapisać gęstość prawdopodobieństwa jako

$$\phi(t_1, \ldots, t_2) \cong \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \frac{t_i^2}{\lambda_i}\right\} .$$

Z powyższego wzoru widać, iż poziomice gęstości są wielowymiarowymi elipsoidami, których półosie są skierowane wzdłuż wektorów własnych $\boldsymbol{\Sigma}$ i mają długości proporcjonalne do $\sqrt{\lambda_i}$.

Powiemy, iż wielowymiarowa zmienna losowa $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ma standardowy wielowymiarowy rozkład normalny jeśli $\boldsymbol{\mu} = \mathbf{0}$ i $\boldsymbol{\Sigma} = \mathbf{1}$. Wówczas

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

Można wykazać, iż jeśli $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ dla $\boldsymbol{\Sigma}$ o niezdegenerowanym widmie to wszystkie rozkłady brzegowe i warunkowe X są rozkładami normalnymi.

1.12 Zbieżność w rachunku prawdopodobieństwa

W rachunku prawdopodobieństwa definiujemy trzy zasadnicze rodzaje zbieżności ciągu zmiennych losowych (X_n).

- Ciąg (X_n) jest zbieżny do X stochastycznie iff

$$\forall \epsilon > 0 : \lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1\,.$$

- Ciąg (X_n) jest zbieżny do X z prawdopodobieństwem 1 iff

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

- Ciąg (X_n) n–wymiarowych zmiennych losowych jest zbieżny do X według dystrybuant iff

$$\forall \mathbf{x} \in \mathbb{R}^n, F_X(\mathbf{x}) \text{ – ciągła w } \mathbf{x} : \lim_{n\to\infty} F_{X_n}(\mathbf{x}) = F_X(\mathbf{x})$$

Pomiędzy tak zdefiniowanymi rodzajami zbieżności zachodzą następujące implikacje:

1. $X_n \to X$ z prawdopodobieństwem 1 $\implies$ $X_n \to X$ stochastycznie
2. $X_n \to X$ stochastycznie $\implies$ $X_n \to X$ według dystrybuant
3. $X_n \to X$ stochastycznie $\implies$ istnieje podciąg (X_{n_k}) zbieżny do X z prawdopodobieństwem 1

1.13 Rozkłady prawdopodobieństwa

1.13.1 Rozkład Bernoulliego

Jeśli zmienna losowa ma wartości w zbiorze $\{x_1, x_2\}$ oraz $P(X = x_1) = p$ i $P(X = x_2) = 1 - p$ (dla $0 \le p \le 1$) to mówimy, że $X \sim \mathrm{Ber}(p)$ (zmienna X ma rozkład Bernoulliego z parametrem p).

1.13.2 Rozkład dwumianowy

Próbą Bernoulliego nazywamy doświadczenie losowe, którego wynik $X \sim \mathrm{Ber}(p)$. Schematem dwumianowym nazywamy n–krotne powtórzenie próby Bernoulliego przy założeniu, iż poszczególne próby są niezależne. Jeśli S będzie zmienną losową

o wartościach w $\mathbb{N}\cup\{0\}$, która opisuje liczbę sukcesów w schemacie dwumianowym długości n. Wówczas

$$P(S=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

i mówimy, że $S\sim \mathrm{Bin}(n,p)$ (zmienna S ma rozkład dwumianowy z parametrami n,p).

1.13.3 Rozkład geometryczny

Rozważamy schemat Bernoulliego o nieskończonej długości. Jeśli T będzie zmienną losową o wartościach w $\mathbb{N}\cup\{0\}$, która opisuje liczbę prób Bernoulliego z paramterem p do momentu uzyskania pierwszego sukcesu to

$$P(T=k)=(1-p)^{k-1}p$$

i mówimy, że $T\sim \mathrm{Geo}(p)$ (zmienna T ma rozkład geometryczny z parametrem p).

1.13.4 Rozkład Poissona

Załóżmy, iż mamy dany skończony przedział czasowy $[0;\tau]$ dla pewnego τ. Niech zmienna losowa N o wartościach w $\mathbb{N}\cup\{0\}$ opisuje liczbę wystąpień pewnego zdarzenia w tym przedziale, przy czym

- zdarzenia występują niezależnie od siebie
- intensywność wystąpień jest stała

to

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

i mówimy, że $N\sim \mathrm{Pois}(\lambda)$ (zmienna N ma rozkład Poissona z parametrem λ). Zachodzi ponadto **twierdzenie Poissona**: niech (S_n) będzie ciągiem takim, że $S_n\sim \mathrm{Bin}(n,p_n)$, gdzie ciąg (p_n) jest taki, iż $\lim_{n\to\infty} np_n=\lambda$, wówczas $\lim_{n\to\infty} P(S_n=k)=\mathrm{e}^{\lambda}\lambda^k/k!$.

1.13.5 Rozkład jednostajny

Jeśli zmienna losowa X o wartościach rzeczywistych ma gęstość prawdopodobieństwa daną przez

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

to mówimy $X\sim\mathcal{U}(a,b)$ (zmienna X ma rozkład jednostajny na odcinku $[a;b]$).

1.13.6 Rozkład wykładniczy

Jeśli zmienna losowa T o wartościach rzeczywistych opisuje prawdopodobieństwo uzyskania pierwszego zdarzenia po czasie x modelowanego przez rozkład Pois(λ) to

$$p(x) = \begin{cases} \lambda \mathrm{e}^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

i mówimy $T \sim \mathrm{Exp}(\lambda)$ (zmienna T ma rozkład wykładniczy z parametrem λ).

1.13.7 Rozkład gamma

Mówimy, że zmienna X o wartościach rzeczywistych ma rozkład gamma z parametrami $p, a > 0$ tj. $X \sim \Gamma(p, a)$ iff gęstość prawdopodobieństwa ma postać

$$p(x) = \begin{cases} \frac{a^p}{\Gamma(p)} x^{p-1} \mathrm{e}^{-ax}, & x > 0 \\ 0, & x \leq 0 \end{cases},$$

gdzie Γ to funkcja Eulera. Parametr p nazywamy parametrem kształtu, a a – parametrem intensywności. Szczególnym przypadkiem rozkładu gamma jest rozkład χ^2 zdefiniowany jako rozkład $\chi^2(n) := \Gamma(n/2, 1/2)$.

2 Elementarz teorii informacji

2.1 Definicja i własności entropii

Mając dany skończony zbiór zdarzeń elementarnych $\{A_1, \ldots, A_n\}$ taki, że wynikiem eksperymentu losowego może być dokładnie jedno z nich oraz prawdopodobieństwa $p_1, \ldots, p_n$, $\sum_i p_i = 1$ każdego z nich powiemy, iż

$$A := \begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

jest *schematem skończonym* (z ang. *finite scheme*). Przykładowo rzut sprawiedliwą, sześcienną kostką do gry jest opisany przez schemat

$$\begin{pmatrix} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$

Zauważmy, że każdy schemat skończony opisuje pewną *niepewność* dotyczącą doświadczenia losowego. Przykładowo jest oczywiste, iż dla schematów

$$\begin{pmatrix} A_1 & A_2 \\ 0.99 & 0.01 \end{pmatrix}, \quad \begin{pmatrix} A_1 & A_2 \\ 0.5 & 0.5 \end{pmatrix}$$

pierwszy z nich opisuje znacznie mniejszą niepewność od drugiego, gdyż prawie z pewnością wynikiem eksperymentu losowego będzie A_1. Wprowadzimy teraz wielkość, która w sensowny sposób mierzy ilość niepewności w danym schemacie skończonym. Wielkością taką jest *entropia Shannona* zdefiniowana dla schematu

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

jako

$$H(A) = H(p_1, p_2, \ldots, p_n) := -\sum_{i=1}^{n} p_i \lg p_i \,,$$

gdzie możemy wybrać dowolną ustaloną podstawę logarytmu oraz stwierdzamy, iż jeśli $p_k = 0$ to $p_k \lg p_k = 0$. Jeśli jako podstawę wybierzemy liczbę 2 to entropię mierzymy w *bitach* tj. 1 bit jest to ilość niepewności zawarta w schemacie skończonym o dwóch jednakowo prawdopodobnych wynikach

$$H = -\log_2 \frac{1}{2} = 1 \,.$$

Przekonamy się teraz, iż tak zdefiniowana miara niepewności ma szereg własności, których spodziewalibyśmy się dla sensownej miary niepewności. Zauważmy wpierw, iż $H(p_1, \ldots, p_n) = 0$ iff dokładnie jedno zdarzenie $A_k \in A$ jest pewne, a pozostałe niemożliwe. Zauważmy dodatkowo, iż z nierówności Jensena mamy dla funkcji wypukłej $\phi(x) = x \lg x$

$$\phi\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i \phi(x_i) \,,$$

dla dowolnych $x_1, \ldots, x_n \in \mathbb{R}$ i $\lambda_1, \ldots, \lambda_n \in [0;1]$, $\sum_i \lambda_i = 1$, skąd

$$\frac{1}{n} \lg \frac{1}{n} \leq \frac{1}{n} \sum_{i=1}^{n} p_i \lg p_i = -\frac{1}{n} H(p_1, \ldots, p_n) \,,$$

czyli

$$H(p_1, \ldots, p_n) \leq -\lg \frac{1}{n} = H(1/n, 1/n, \ldots, 1/n) \,,$$

czyli niepewność zawarta w danym schemacie skończonym jest mniejsza lub równa od niepewności zawartej w analogicznym schemacie, w którym wszystkie wyniki są jednakowo prawdopodobne.

Załóżmy teraz, że mamy dwa niezależne schematy skończone

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}, \quad B = \begin{pmatrix} B_1 & B_2 & \cdots & B_m \\ q_1 & q_2 & \cdots & q_m \end{pmatrix}$$

takie, że dla każdej pary zdarzeń A_i, B_j prawdopodobieństwo wystąpienia zdarzenia A_iB_j wynosi p_iq_j. Zbiór zdarzeń A_iB_j z prawdopodobieństwami $r_{ij} = p_iq_j$ reprezentuje nowy schemat skończony AB. Wówczas

$$\begin{aligned} -H(AB) &= \sum_{i=1}^{n}\sum_{j=1}^{m} r_{ij} \lg r_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{m} p_iq_j \left(\lg p_i + \lg q_j\right) \\ &= \sum_{i=1}^{n} p_i \lg p_i + \sum_{j=1}^{m} q_j \lg q_j = -H(A) - H(B)\,, \end{aligned}$$

skąd

$$H(AB) = H(A) + H(B)\,.$$

Rozważmy teraz przypadek gdy schematy A, B są zależne. Przez q_{ij} oznaczmy prawdopodobieństwo zajścia zdarzenia B_j pod warunkiem zdarzenia A_i tj. $q_{ij} = p(B_j \mid A_i)$. Schemat AB jest teraz opisany prawdopodobieństwami $r_{ij} = p_iq_{ij}$ zatem

$$-H(AB) = \sum_{i=1}^{n}\sum_{j=1}^{m} p_iq_{ij} \left(\lg p_i + \lg q_{ij}\right) = -H(A) + \sum_{i=1}^{n} p_i \sum_{j=1}^{m} q_{ij} \lg q_{ij}$$

gdyż $\sum_j q_{ij} = 1$ (prawdopodobieństwo zajścia dowolnego zdarzenia z B pod warunkiem wystąpienia zdarzenia A_i wynosi 1), natomiast wielkość $-\sum_{j=1}^{m} q_{ij} \lg q_{ij}$ jest warunkową entropią schematu B pod warunkiem zajścia zdarzenia A_i, co oznaczymy jako $H(B \mid A = A_i)$

$$H(AB) = H(A) + \sum_{i=1}^{n} p_i H(B \mid A = A_i)\,.$$

Ostatni człon jest w takim razie wartością oczekiwaną wielkości $H(B)$ w schemacie A, co oznaczymy jako $H(B \mid A)$. Mamy w takim razie

$$H(AB) = H(A) + H(B \mid A)\,.$$

Z nierówności Jensena można dodatkowo pokazać, że zachodzi $H(B \mid A) \leq H(B)$.

2.2 Entropia względna

Dla dwóch ciągłych rozkładów prawdopodobieństwa $p(\mathbf{x})$, $q(\mathbf{x})$ definiujemy ich entropię względną (nazywaną również *Kullback-Leibler (KL) divergence*) jako

$$\mathbb{D}_{\mathrm{KL}}(p, q) = \int\limits_{\mathbb{R}^n} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}\, \mathrm{d}^n\mathbf{x}\ ,$$

która określa podobieństwo między dwoma rozkładami prawdopodobieństwa tj. dla ustalonego rozkładu p dla wszystkich q zachodzi $\mathbb{D}_{\mathrm{KL}}(p,q) \geq 0$, przy czym równość zachodzi iff $p = q$ (ponownie nierówność Jensena).

Rozważmy teraz rozkład łączny $p(\mathbf{x}, \mathbf{y})$. Jeśli zmienne losowe $\mathbf{x}, \mathbf{y}$ są niezależne to $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$. Jeśli zmienne nie są niezależne to możemy określić stopień ich zależności właśnie poprzez entropię względną między rozkładem łącznym $p(\mathbf{x}, \mathbf{y})$, a rozkładem faktoryzowanym $p(\mathbf{x})p(\mathbf{y})$. Wielkość taką nazywamy informacją wzajemną (z ang/ *mutual information*)

$$\mathbb{I}(\mathbf{x}, \mathbf{y}) = \mathbb{D}_{\mathrm{KL}}(p(\mathbf{x}, \mathbf{y}), p(\mathbf{x})p(\mathbf{y})) = \int\limits_{\mathbb{R}^n \times \mathbb{R}^m} p(\mathbf{x}, \mathbf{y}) \log \left\{ \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})} \right\} \mathrm{d}^n\mathbf{x}\, \mathrm{d}^m\mathbf{y} \ .$$

3 Statystyka

3.1 Wnioskowanie statystyczne

Modelem statystycznym nazwiemy parę $(\chi, \mathcal{P})$, gdzie $\mathcal{P}$ jest rodziną rozkładów prawdopodobieństwa w zbiorze χ, przy czym będziemy zakładać $\chi = \mathbb{R}^n$

$$\mathcal{P} := \{p(\mathbf{x} \mid \theta) \mid \theta \in \Theta\} \ ,$$

gdzie Θ jest zbiorem parametrów modelu $\mathcal{P}$. Prostą próbą losową w modelu $\mathcal{P}$ nazwiemy ciąg niezależnych zmiennych losowych $X_1, \ldots, X_n$ o wartościach w $\mathbb{R}^n$ i pochodzących z tego samego rozkładu $p(\mathbf{x} \mid \theta) \in \mathcal{P}$ (w angielskiej terminologii taki ciąg zmiennych losowych nazwiemy *i.i.d.* tj. *independent and identically distributed*). Statystyką z kolei nazwiemy zmienną losową T będącą funkcją prostej próby losowej tj. $T = T(X_1, \ldots, X_n)$. Być może najważniejszym przykładem statystyki jest średnia oznaczana jako $\overline{X}$

$$\overline{X}(X_1, \ldots, X_n) := \frac{X_1 + \ldots + X_n}{n} \, .$$

Wartość oczekiwana statystyki średniej $\overline{X}(X_1, \ldots, X_n)$ dla X_i z rozkładu $X \sim \mathcal{D}$ o gęstości p wynosi

$$\mathbb{E}[\overline{X}] = \int \cdots \int \frac{1}{n} \left(\sum_{i=1}^{n} X_i \right) p(X_1) \cdots p(X_n) \, \mathrm{d}X_1 \ldots \mathrm{d}X_n = \mathbb{E}[X] \, .$$

Wariancja statystyki średniej wynosi z kolei

$$\begin{aligned}
&\mathrm{Var}[\overline{X}] = \mathbb{E}[\overline{X}^2] - \mathbb{E}[\overline{X}]^2 \\
&= \int \cdots \int \frac{1}{n^2} \left(\sum_{i=1}^{n} X_i^2 + \underbrace{\sum_{i \neq j} X_i X_j}_{n(n-1)} \right) p(X_1) \cdots p(X_n) \, \mathrm{d}X_1 \ldots \mathrm{d}X_n - \mathbb{E}[X]^2 \\
&= \frac{1}{n}\mathbb{E}[X^2] + \frac{n(n-1)}{n^2}\mathbb{E}[X]^2 - \mathbb{E}[X]^2 = \frac{1}{n}\left[\mathbb{E}[X^2] - \mathbb{E}[X]^2\right] = \frac{1}{n}\mathrm{Var}[X] \, .
\end{aligned}$$

3.2 Twierdzenie Gliwenki–Cantelliego

Dystrybuantą empiryczną nazywa się funkcję

$$\hat{F}(x) = \frac{1}{N} \# \left\{ i \in \{1, \ldots, N\} \mid x_i \leq x \right\} ,$$

gdzie $\{x_1, \ldots, x_N\}$ jest realizacją prostej próby losowej. Twierdzenie Gliwenki–Cantelliego stwierdza, iż jeśli $F(x)$ jest dystrybuantą pewnego rozkładu prawdopodobieństwa to

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)| \to 0 \, , \text{przy } n \to \infty .$$

3.3 Silne prawo wielkich liczb

Niech (X_n) będzie ciągiem zmiennych losowych i.i.d. z pewnego rozkładu $X \sim \mathcal{D}$. Przez $(\overline{X}_n)$ oznaczmy ciąg średnich częściowych tj.

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

Wówczas zachodzi silne prawo wielkich liczb

$$P\left(\lim_{n \to \infty} \overline{X}_n = \mathbb{E}[X] \right) = 1 \, ,$$

czyli średnia próbek zbiega do wartości oczekiwanej z prawdopodobieństwem 1.

Silne prawo wielkich liczb daje nam potężne narzędzie do szacowania wartości oczekiwanych, gdyż możemy je przybliżać średnią z dużej liczby próbek losowych, a dokładność tego przybliżenia zależy jedynie od liczby próbek i wariancji X. Jeśli X jest zmienną wielowymiarową to dokładność przybliżenia nie zależy wprost od liczby wymiarów i unikamy tzw. *curse of dimensionality*.

3.4 Centralne Twierdzenie Graniczne

Niech (X_n) będzie ciągiem k–wymiarowych zmiennych losowych i.i.d. z dowolnego rozkładu $X \sim \mathcal{D}$ o wartości oczekiwanej $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$ i odwracalnej macierzy kowariancji $\boldsymbol{\Sigma}$. Oznaczając przez $(\overline{X}_n)$ ciąg średnich częściowych ciągu (X_n) zachodzi

$$\sqrt{n}\left(\overline{X}_n - \boldsymbol{\mu}\right) \to Z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})\,.$$

Oznacza to, iż dla ciągu $X_1, \ldots, X_n$ zmiennych losowych i.i.d. z praktycznie dowolnego rozkładu $X \sim \mathcal{D}$ dla odpowiednio dużych n średnią z próbek możemy traktować jako zmienną losową o rozkładzie normalnym $\mathcal{N}(\boldsymbol{\mu}, n^{-1/2}\boldsymbol{\Sigma})$.

3.5 Estymatory punktowe MLE i MAP

Rozważamy model statystyczny $\mathcal{P} = \{p(\mathbf{x} \mid \theta) \mid \theta \in \Theta\}$. Estymatorem parametru θ nazwiemy statystykę $\hat{\theta}(X_1, \ldots, X_n)$ służącą do oszacowania wartości tego parametru. Wartość tej statystki dla konkretnej realizacji prostej próby losowej $\hat{\theta}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ nazwiemy estymatą parametru θ. Dodatkowo definiujemy obciążenie (z ang. *bias*) estymatora jako wielkość

$$\mathbb{B}[\hat{\theta}] := \mathbb{E}[\hat{\theta}] - \theta\,.$$

Zasadniczo będą nas interesować dwa rodzaje estymat: MLE i MAP. W przypadku estymaty MLE (z ang. *Maximum Likelihood Estimate*) definiujemy funkcję wiarygodności (*likelihood*) dla modelu $\mathcal{P} = \{p(\mathbf{x} \mid \theta) \mid \theta \in \Theta\}$ i realizacji prostej próby losowej (którą nazwiemy również danymi lub obserwacjami) $D = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ jako

$$p(D \mid \theta) = \prod_{i=1}^{n} p(\mathbf{x}_i \mid \theta)\,.$$

Estymatą MLE nazywamy taką wartość parametru $\theta_{\mathrm{MLE}} \in \Theta$, że

$$p(D \mid \theta_{\mathrm{MLE}}) = \max_{\theta \in \Theta} p(D \mid \theta)\,.$$

Ponieważ znajdywanie maksimum funkcji będącej iloczynem nie jest zadaniem przyjemnym (chociażby obliczanie pochodnych iloczynu funkcji jest trudniejsze od sumy), więc wprowadzamy zanegowaną logarytmiczną funkcję wiarygodności

$$\ell(D \mid \theta) = -\log p(D \mid \theta) = -\sum_{i=1}^{n} \log p(\mathbf{x}_i \mid \theta)\,,$$

wówczas ze względu na fakt, iż funkcja $\log x$ jest ściśle rosnąca estymatę MLE możemy równoważnie wyznaczyć jako

$$\ell(D \mid \theta_{\mathrm{MLE}}) = \min_{\theta \in \Theta} \ell(D \mid \theta)\,.$$

Funkcję ℓ będziemy również nazywać funkcją kosztu.

W przypadku estymaty MAP (z ang. *Maximum a posteriori estimate*) wprowadzamy gęstość rozkładu a posteriori jako

$$p(\theta \mid D) = \frac{1}{Z} p(D \mid \theta)\pi(\theta)\,,$$

gdzie Z jest stałą wynikającą z warunku unormowania, a $\pi(\theta)$ to gęstość prawdopodobieństwa opisująca rozkład a priori parametru θ. Estymatą MAP nazywamy taką wartość parametru $\theta_{\mathrm{MAP}} \in \Theta$, że

$$p(\theta_{\mathrm{MAP}} \mid D) = \max_{\theta \in \Theta} p(\theta \mid D)\,.$$

Zauważmy przy tym iż liczba Z nie jest nam potrzebna, gdyż wystarczy zmaksymalizować licznik tj.

$$\theta_{\mathrm{MAP}} = \arg\max_{\theta \in \Theta} p(D \mid \theta)\pi(\theta)\,.$$

4 Probabilistyczne uczenie maszynowe

4.1 Wnioskowanie Bayesowskie

Zajmiemy się teraz wnioskowaniem opartym na twierdzeniu Bayesa. Rozpatrujemy model statystyczny $\mathcal{P} = \{p(\mathbf{x} \mid \theta) \mid \theta \in \Theta\}$. Załóżmy, iż mamy obserwacje $D = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, wówczas twierdzenie Bayesa możemy zapisać jako

$$p(\theta \mid D) = \frac{p(D \mid \theta)\pi(\theta)}{p_D(D)} = \frac{p(D \mid \theta)\pi(\theta)}{\int\limits_{\Theta} p(D \mid \theta)\pi(\theta)\,\mathrm{d}\theta}\,,$$

gdzie $p(\theta \mid D)$ nazywamy rozkładem a posteriori (posteriorem), $p(D \mid \theta)$ – wiarygodnością (likelihood), a $\pi(\theta)$ – rozkładem a priori (priorem).

Całe wnioskowanie Bayesowskie opiera się na wyznaczeniu rozkładu a posteriori, który wyraża całą naszą wiedzę o estymowanym parametrze θ. Na podstawie tego rozkładu możemy wyznaczyć estymatę punktową MAP maksymalizującą gęstość prawdopodobieństwa a posteriori, jak również niepewność związaną z wyznaczeniem tej estymaty np. poprzez wyznaczenie przedziału wiarygodności $C_{1-\alpha}(\theta \mid D) = [\theta_l; \theta_u]$ takiego, że

$$P(\theta \in [\theta_l; \theta_u] \mid D) = 1 - \alpha\,,$$

dla ustalonego $0 < \alpha < 1$. Możemy również skonstruować rozkład predykcyjny (z ang. *posterior predictive distribution*) określający prawdopodobieństwo zaobserwowania nowej obserwacji $\mathbf{x}$

$$p(\mathbf{x} \mid D) = \int\limits_{\Theta} p(\mathbf{x} \mid \theta)p(\theta \mid D)\,\mathrm{d}\theta\,.$$

Znając rozkład a posteriori estymowanego parametru θ 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 kosztu (z ang. *loss/cost function*) $L(\theta, \hat{\theta})$ po wszystkich estymatorach $\hat{\theta}$

$$\theta_{\text{Bayes}} = \arg\min_{\hat{\theta}} \int_{\Theta} L(\theta, \hat{\theta}) p(\theta \mid D)\, \mathrm{d}\theta \ .$$

Całkę w powyższym wzorze nazywa się również funkcją ryzyka (z ang. *risk function*) $R(\hat{\theta})$, która określa oczekiwaną stratę spowodowaną wykorzystaniem danego estymatora parametru θ. W przypadku gdy funkcja kosztu ma postać błędu kwadratowego (L2)

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

funkcję ryzyka możemy zapisać jako

$$\begin{aligned} R(\hat{\theta}) &= \int_{\Theta} \theta^2 p(\theta \mid D)\, \mathrm{d}\theta - 2\hat{\theta} \int_{\Theta} \theta p(\theta \mid D)\, \mathrm{d}\theta + \hat{\theta}^2 \\ &= \mathrm{Var}[\theta \mid D] + \mathbb{E}[\theta \mid D]^2 - 2\hat{\theta}\mathbb{E}[\theta \mid D] + \hat{\theta}^2 \\ &= \mathrm{Var}[\theta \mid D] + \left(\mathbb{E}[\theta \mid D] - \hat{\theta}\right)^2 . \end{aligned}$$

4.2 Bayesowski wybór modeli

Załóżmy, iż mamy rodzinę $\mathcal{M}$ modeli statystycznych (może to być zbiór dyskretny lub zbiór modeli indeksowanych ciągłym, wielowymiarowym parametrem $\boldsymbol{\lambda}$). Naszym zadaniem jest wybór najbardziej prawdopodobnego modelu dla danych D. Możemy na to zadanie patrzeć jako zadanie z teorii decyzji: dla danej funkcji kosztu $L(M, M^*)$ i rozkładu a posteriori nad modelami $p(M \mid D)$ chcemy wybrać model, który minimalizuje ryzyko $\mathbb{E}[L(M, M^*)]$. Jeśli jako koszt wybierzemy tzw. *0–1 loss* tj.

$$L(M, M^*) = \begin{cases} 0 & \text{, jeśli } M = M^* \\ 1 & \text{, w.p.p.} \end{cases}$$

to

$$\mathbb{E}[L(M, M^*)] = 1 - p(M^* \mid D)$$

i wybieramy model M o największym prawdopodobieństwie (estymata MAP). Pozostaje tylko wyznaczenie $p(M \mid D)$

$$p(M \mid D) = \frac{p(D \mid M)\pi(M)}{\sum_{M \in \mathcal{M}} p(D \mid M)\pi(M)} \, .$$

Jeśli jako prior przyjmiemy rozkład jednostajny $\pi(M) = |\mathcal{M}|^{-1}$ to estymata MAP sprowadza się do MLE czyli szukamy modelu

$$M^* = \arg\max_{M \in \mathcal{M}} p(D \mid M)\,.$$

Jeśli przez θ_M oznaczymy parametry modelu M to

$$p(D \mid M) = \int\limits_{\Theta_M} p(D \mid \theta_M)\pi(\theta_M)\,\mathrm{d}\theta_M\ .$$

Powyższą wielkość nazywamy wiarygodnością brzegową (z ang. *marginal likelihood*) lub *model evidence*.

4.3 Estymator jądrowy gęstości (KDE)

Załóżmy, że mamy zbiór obserwacji iid $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ taki, że $\mathbf{x}_i \sim \mathcal{D}$ dla pewnego n–wymiarowego ciągłego rozkładu prawdopodobieństwa $\mathcal{D}$ z nieznaną gęstością prawdopodobieństwa $p(\mathbf{x})$. Chcemy znaleźć estymator $\hat{p}(\mathbf{x})$ tej funkcji. Estymatorem jądrowym gęstości funkcji p (z ang. *kernel density estimator*) nazywamy funkcję

$$\hat{p}(\mathbf{x}) := \frac{1}{mh^n} \sum_{i=1}^{m} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)\,,$$

gdzie $h \in \mathbb{R}$ jest pewnym hiperparametrem zwanym *bandwidth*, a $K : \mathbb{R}^n \mapsto [0; \infty)$ to tzw. funkcja jądrowa będąca parzystą funkcją posiadającą w 0 maksimum globalne oraz spełniającą warunek unormowania

$$\int\limits_{\mathbb{R}^n} K(\mathbf{x})\,\mathrm{d}^n\mathbf{x} = 1\,.$$

Ze statystycznego punktu widzenia, postać jądra nie ma istotnego znaczenia i wybór funkcji K może być arbitralny, uwzględniający przede wszystkim pożądane własności otrzymanego estymatora, na przykład klasę jego regularności (ciągłość, różniczkowalność itp.). W przypadku jednowymiarowym jako funkcję K przyjmuje się klasyczne postacie gęstości rozkładów probabilistycznych, na przykład gęstość rozkładu normalnego. W przypadku wielowymiarowym stosuje się tzw. jądro radialne tj. dla jądra jednowymiarowego K wielowymiarowe jądro radialne definiujemy jako

$$K(\mathbf{x}) = K(\|\mathbf{x}\|)$$

dla pewnej normy (typowo normy euklidesowej) $\|\cdot\|$.

4.4 Modele Gaussowskie

Jak już wspomnieliśmy w przypadku gdy zmienna losowa ma wielowymiarowy rozkład normalny $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ wszystkie rozkłady brzegowe i warunkowe są również rozkładami normalnymi. W szczególnym przypadku gdy zmienne k i $n-k$ –wymiarowe $\mathbf{x}$ i $\mathbf{y}$ mają łącznie rozkład normalny

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

gdzie

$$\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_{\mathbf{x}}\\ \boldsymbol{\mu}_{\mathbf{y}}\end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}}\\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{bmatrix}$$

można pokazać iż

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

gdzie

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

4.5 Liniowe modele Gaussowskie

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

$$\mathbf{y} \mid \mathbf{x} \sim \mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_{\mathbf{y}}),$$

gdzie $\mathbf{A}$ jest macierzą wymiaru $n \times k$. Jako prior na parametr $\mathbf{x}$ przyjmiemy również rozkład normalny o pewnych zadanych parametrach $\boldsymbol{\mu}_{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}}$ (taki wybór rozkładu a priori nazywamy rozkładem sprzężonym do wiarygodności)

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

Wówczas łatwo pokazać, iż rozkład a posteriori jest rozkładem normalnym

$$\mathbf{x} \mid \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}})$$

z parametrami

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}} &= \left[\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} + \mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{A}\right]^{-1}\\ \boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} &= \boldsymbol{\Sigma}_{\mathbf{x}|\mathbf{y}}\left[\mathbf{A}^{\top}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\mu}_{\mathbf{x}}\right]\end{aligned}.$$

Załóżmy teraz, iż mamy ciąg obserwacji $(\mathbf{y}_1, \ldots, \mathbf{y}_m)$. Wnioskowanie Bayesowskie możemy wówczas stosować iteracyjnie tzn. na początku dla 0 obserwacji rozkład estymowanego parametru jest opisany przez prior $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Po zaobserwowaniu jednego $\mathbf{y}_1$ aktualizujemy nasze przekonania co do parametru $\mathbf{x}$ zgodnie z powyższym wzorem i otrzymujemy rozkład normalny o parametrach

$$\begin{aligned}\boldsymbol{\Sigma}_1 &= \left[\boldsymbol{\Sigma}_0^{-1} + \mathbf{A}^\top \boldsymbol{\Sigma}_\mathbf{y}^{-1} \mathbf{A}\right]^{-1} \\ \boldsymbol{\mu}_1 &= \boldsymbol{\Sigma}_1 \left[\mathbf{A}^\top \boldsymbol{\Sigma}_\mathbf{y}^{-1} (\mathbf{y}_1 - \mathbf{b}) + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0\right]\end{aligned} .$$

Po zaobserwowaniu kolejnego $\mathbf{y}_2$ ponownie wykorzystujemy powyższe wzory ale jako prior wykorzystując rozkład w poprzedniej iteracji. W ogólności możemy zapisać wzór rekurencyjny na $m+1$ rozkład jako

$$\begin{aligned}\boldsymbol{\Sigma}_{m+1} &= \left[\boldsymbol{\Sigma}_m^{-1} + \mathbf{A}^\top \boldsymbol{\Sigma}_\mathbf{y}^{-1} \mathbf{A}\right]^{-1} \\ \boldsymbol{\mu}_{m+1} &= \boldsymbol{\Sigma}_{m+1} \left[\mathbf{A}^\top \boldsymbol{\Sigma}_\mathbf{y}^{-1} (\mathbf{y}_{m+1} - \mathbf{b}) + \boldsymbol{\Sigma}_m^{-1} \boldsymbol{\mu}_m\right]\end{aligned} ,$$

skąd możemy od razu podać wzór na parametry m–tego rozkładu

$$\begin{aligned}\boldsymbol{\Sigma}_m &= \left[\boldsymbol{\Sigma}_0^{-1} + m\mathbf{A}^\top \boldsymbol{\Sigma}_\mathbf{y}^{-1} \mathbf{A}\right]^{-1} \\ \boldsymbol{\mu}_m &= \boldsymbol{\Sigma}_m \left[\mathbf{A}^\top \boldsymbol{\Sigma}_\mathbf{y}^{-1} \left(\sum_{i=1}^m \mathbf{y}_i - m\mathbf{b}\right) + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0\right]\end{aligned} .$$

Taki sam wynik można by uzyskać rozpatrując łączny rozkład a posteriori dla obserwacji $D = (\mathbf{y}_1, \ldots, \mathbf{y}_m)$ tj.

$$p(\mathbf{x} \mid D) \cong \pi(\mathbf{x}) \prod_{i=1}^m p(\mathbf{y}_i \mid \mathbf{x}) \cong$$

$$\exp\left\{-\frac{1}{2}\left[(\mathbf{x} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\mathbf{x} - \boldsymbol{\mu}_0) + \sum_{i=1}^m (\mathbf{y}_i - \mathbf{A}\mathbf{x} - \mathbf{b})^\top \boldsymbol{\Sigma}_\mathbf{y}^{-1} (\mathbf{y}_i - \mathbf{A}\mathbf{x} - \mathbf{b})\right]\right\} .$$

4.6 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, który kontrolujemy tj. zakładamy, iż wektor $\mathbf{x}$ dla danego pomiaru y znamy dokładnie. Dodatkowo zakładamy, iż y zależy liniowo od $\mathbf{x}$ tj.

$$y = \mathbf{w}^\top \mathbf{x} + \epsilon \,,$$

gdzie $\epsilon \sim \mathcal{N}(0, \sigma^2)$ dla znanego σ jest tzw. błędem losowym, a $\mathbf{w}$ jest estymowanym przez nas parametrem. Możemy zatem zapisać

$$y \mid \mathbf{w} \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2) \,.$$

Powiedzmy, iż zaobserwowaliśmy ciąg obserwacji $D = (y_1, \ldots, y_m)$ dla zadanych (lub dokładnie znanych) przez nas $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$. Wiarygodność ma zatem postać

$$p(D \mid \mathbf{w}) \cong \prod_{i=1}^{m} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - \mathbf{w}^\top \mathbf{x}_i\right)^2\right\} .$$

W przypadku regresji liniowej zamiast pełnego wnioskowania Bayesowskiego o parametrze $\mathbf{w}$ często stosuje się prostsze podejście polegające na znalezieniu estymaty punktowej MLE. Zanegowana logarytmiczna funkcja wiarygodności ma postać

$$\ell(D \mid \mathbf{w}) = \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \text{const.}$$

Człon stały możemy oczywiście pominąć i zapisać

$$\ell(D \mid \mathbf{w}) \cong \sum_{i=1}^{m} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) ,$$

gdzie

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} , \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_m^\top \end{bmatrix} .$$

Ponieważ otrzymana funkcja ℓ ma postać formy kwadratowej, więc problem optymalizacyjny polegający na znalezieniu minimum ℓ nazywa się metodą najmniejszych kwadratów (z ang. *OLS – Ordinary Least Squares*). Aby wyznaczyć estymatę $\mathbf{w}_{\text{MLE}}$ musimy rozwiązać równanie

$$\frac{\partial \ell}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left[\mathbf{y}^\top \mathbf{y} + \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{y}^\top \mathbf{X}\mathbf{w}\right] = \mathbf{0} ,$$

skąd

$$2\mathbf{X}^\top \mathbf{X}\mathbf{w} - 2\mathbf{X}^\top \mathbf{y} = \mathbf{0} ,$$

zatem

$$\mathbf{w}_{\text{MLE}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} .$$

Pełniejszą informację o parametrze $\mathbf{w}$ możemy uzyskać rozpatrując rozkład a posteriori $p(\mathbf{w} \mid D)$. Jeśli jako prior przyjmiemy rozkład normalny z pewnymi parametrami $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$ to zauważmy, iż otrzymujemy instancję liniowego modelu Gaussowskiego

$$\begin{aligned} \mathbf{y} \mid \mathbf{w} &\sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{1}) \\ \mathbf{w} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \end{aligned} ,$$

skąd rozkład a posteriori jest rozkładem normalnym

$$\mathbf{w} \mid \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

o parametrach

$$\begin{aligned} \boldsymbol{\Sigma}_m &= \left[\boldsymbol{\Sigma}_0^{-1} + \sigma^{-2}\mathbf{X}^\top\mathbf{X}\right]^{-1} \\ \boldsymbol{\mu}_m &= \boldsymbol{\Sigma}_m\left[\sigma^{-2}\mathbf{X}^\top\mathbf{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right] \end{aligned} .$$

W powyższych wzorach nazwy parametrów nie są przykładowe: po zaobserwowaniu 0 przykładów rozkład parametru $\mathbf{w}$ jest rozkładem a priori $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$; po zaobserwowaniu po jednej wartości y_i w m zadanych (znanych dokładnie) punktach $\mathbf{x}_i$ otrzymujemy rozkład a posteriori $\mathcal{N}(\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$. Gdybyśmy w każdym z m punktów $\mathbf{x}_i$ dokonywali pomiaru y_i s–krotnie to wtedy wykorzystując wzory wyprowadzone przy iteracyjnym stosowaniu wnioskowania w liniowym modelu Gaussowskim otrzymujemy rozkład normalny o parametrach

$$\begin{aligned} \boldsymbol{\Sigma}_{m;s} &= \left[\boldsymbol{\Sigma}_0^{-1} + \frac{s}{\sigma^2}\mathbf{X}^\top\mathbf{X}\right]^{-1} \\ \boldsymbol{\mu}_{m;s} &= \boldsymbol{\Sigma}_{m;s}\left[\sigma^{-2}\mathbf{X}^\top\sum_{i=1}^{s}\mathbf{y}_i + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right] \end{aligned} .$$

Rozkład predykcyjny dla nowej obserwacji y poczynionej w punkcie $\mathbf{x}$ jest dany przez

$$p(y \mid \mathbf{y}) = \int_{\mathbb{R}^n} p(y \mid \mathbf{w}) p(\mathbf{w} \mid \mathbf{y}) \mathrm{d}^n\mathbf{w} \,.$$

Nietrudno zauważyć, iż będzie to rozkład normalny o parametrach

$$\begin{aligned} \mu_{y|\mathbf{y}} = \mathbb{E}[y \mid \mathbf{y}] &= \int_{\mathbb{R}} yp(y \mid \mathbf{y})\,\mathrm{d}y = \int_{\mathbb{R}^n} \mathrm{d}^n\mathbf{w}\, p(\mathbf{w} \mid \mathbf{y}) \int_{\mathbb{R}} \mathrm{d}y\, yp(y \mid \mathbf{w}) \\ &= \int_{\mathbb{R}^n} \mathrm{d}^n\mathbf{w}\, p(\mathbf{w} \mid \mathbf{y})\mathbf{x}^\top\mathbf{w} = \mathbf{x}^\top\boldsymbol{\mu}_m \,. \end{aligned}$$

oraz

$$\begin{aligned}
\sigma^2_{y|\mathbf{y}} &= \mathbb{E}\left[(y-\mu_{y|\mathbf{y}})^2 \mid \mathbf{y}\right] = \int\limits_{\mathbb{R}^n} \mathrm{d}^n\mathbf{w}\, p(\mathbf{w}\mid\mathbf{y}) \int\limits_{\mathbb{R}} \mathrm{d}y\,(y-\mu_{y|\mathbf{y}})^2 p(y\mid\mathbf{w}) \\
&= \int\limits_{\mathbb{R}^n} \mathrm{d}^n\mathbf{w}\, p(\mathbf{w}\mid\mathbf{y}) \int\limits_{\mathbb{R}} \mathrm{d}y\left(y^2+\mu^2_{y|\mathbf{y}}-2\mu_{y|\mathbf{y}}y\right) p(y\mid\mathbf{w}) \\
&= \int\limits_{\mathbb{R}^n} \mathrm{d}^n\mathbf{w}\, p(\mathbf{w}\mid\mathbf{y}) \left(\sigma^2+(\mathbf{x}^\top\mathbf{w})^2+\mu^2_{y|\mathbf{y}}-2\mu_{y|\mathbf{y}}\mathbf{x}^\top\mathbf{w}\right) \\
&= \sigma^2 + \int\limits_{\mathbb{R}^n} \mathrm{d}^n\mathbf{w}\, p(\mathbf{w}\mid\mathbf{y}) \left(\mathbf{x}^\top\mathbf{w}-\mathbf{x}^\top\boldsymbol{\mu}_m\right)^2 \\
&= \sigma^2 + \mathbf{x}^\top\mathbb{E}[(\mathbf{w}-\boldsymbol{\mu}_m)(\mathbf{w}-\boldsymbol{\mu}_m)^\top \mid \mathbf{y}]\mathbf{x} = \sigma^2+\mathbf{x}^\top\boldsymbol{\Sigma}_m\mathbf{x}\,.
\end{aligned}$$

Powyżej skorzystaliśmy ze znanego faktu, iż dla jednowymiarowej zmiennej losowej zachodzi $\sigma^2 = \mathbb{E}[(X-\mu_X)^2] = \mathbb{E}[X^2]-\mu_X^2$, skąd $\mathbb{E}[X^2] = \sigma^2+\mu_X^2$. Podsumowując rozkład predykcyjny ma postać

$$y \mid \mathbf{y} \sim \mathcal{N}(\mathbf{x}^\top\boldsymbol{\mu}_m, \sigma^2+\mathbf{x}^\top\boldsymbol{\Sigma}_m\mathbf{x})\,.$$

4.7 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 postaci

$$\gamma\|\theta\|$$

dla pewnej normy $\|\cdot\|$ i hiper-parametru γ określającego siłę regularyzacji. W kontekście Bayesowskim regularyzację można również rozumieć jako pewną niechęć ("tłumienie", zachowawczość) modelu do zmiany rozkładu a priori estymowanego parametru po pojawieniu się kolejnych obserwacji.

Przykładowo jeśli w zagadnieniu Bayesowskiej regresji liniowej jako prior przyjmiemy rozkład normalny

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{1})$$

to rozkład a posteriori jest rozkładem normalnym o parametrach

$$\begin{aligned}
\boldsymbol{\Sigma}_m &= \sigma^2\left[\gamma\mathbf{1}+\mathbf{X}^\top\mathbf{X}\right]^{-1} \\
\boldsymbol{\mu}_m &= \left[\gamma\mathbf{1}+\mathbf{X}^\top\mathbf{X}\right]^{-1}\mathbf{X}^\top\mathbf{y}
\end{aligned}\quad,$$

gdzie $\gamma = \sigma^2/\tau^2$ jest hiper-parametrem określającym siłę regularyzacji. Zauważmy, że im większa jest wartość γ (mniejsza niepewność związana z rozkładem a priori) tym drugi człon w nawiasie staje się mniej istotny. Taki sam wynik możemy uzyskać metodą OLS jeśli do funkcji kosztu dodamy człon regularyzujący dla zwykłej normy euklidesowej. Zagadnienie minimalizacji funkcji kosztu będącej formą kwadratową z dodanym członem regularyzującym nazywamy również regresją grzbietową.

4.8 Procesy Gaussowskie

Jak już wspomnieliśmy macierz kowariancji n–wymiarowej zmiennej losowej $\mathbf{x}$ o wartości oczekiwanej $\boldsymbol{\mu}$ jest zdefiniowana jako

$$\boldsymbol{\Sigma} = \mathbb{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\right] .$$

Pokazaliśmy również, iż macierz ta jest nieujemnie określona. Dodatkowo pokażemy, iż dla każdej nieujemnie określonej macierzy symetrycznej $\mathbf{K}$ wymiaru $n \times n$ istnieje n–wymiarowa zmienna losowa o wielowymiarowym rozkładzie normalnym dla której $\mathbf{K}$ jest macierzą kowariancji. Istotnie dla każdej nieujemnie określonej macierzy symetrycznej istnieje macierz $\mathbf{L}$ taka, że

$$\mathbf{K} = \mathbf{L}\mathbf{L}^\top ,$$

jest to tzw. dekompozycja Choleskiego. Niech $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$, wówczas zmienna losowa $\mathbf{Lz}$ ma rozkład o zerowej wartości oczekiwanej i macierzy kowariancji

$$\mathbb{E}\left[(\mathbf{Lz})(\mathbf{Lz})^\top\right] = \mathbb{E}\left[\mathbf{L}\mathbf{z}\mathbf{z}^\top\mathbf{L}^\top\right] = \mathbf{L}\mathbb{E}[\mathbf{z}\mathbf{z}^\top]\mathbf{L}^\top = \mathbf{L}\mathbf{1}\mathbf{L}^\top = \mathbf{K} .$$

Powyższe własności wskazują, iż macierze kowariancji można w pewnym sensie utożsamiać z nieujemnie określonymi macierzami symetrycznymi.

Zdefiniujemy teraz funkcję kowariancji $k : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ taką, że $\forall m \in \mathbb{N}$: $\forall X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathbb{R}^n$ macierz

$$k(X, X) = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & k(\mathbf{x}_1, \mathbf{x}_2) & \cdots & k(\mathbf{x}_1, \mathbf{x}_m) \\ k(\mathbf{x}_2, \mathbf{x}_1) & k(\mathbf{x}_2, \mathbf{x}_2) & \cdots & k(\mathbf{x}_2, \mathbf{x}_m) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_m, \mathbf{x}_1) & k(\mathbf{x}_m, \mathbf{x}_2) & \cdots & k(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix}$$

jest dodatnio określoną macierzą symetryczną. Funkcję k nazywamy również jądrem dodatnio określonym (z ang. *positive definite kernel*) lub jądrem Mercera. Dla dwóch zbiorów punktów $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathbb{R}^n$ i $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_s\} \subset \mathbb{R}^n$ i

funkcji kowariancji k wprowadzimy oznaczenie

$$k(X,Y) := \begin{bmatrix} k(\mathbf{x}_1,\mathbf{y}_1) & k(\mathbf{x}_1,\mathbf{y}_2) & \cdots & k(\mathbf{x}_1,\mathbf{y}_s) \\ k(\mathbf{x}_2,\mathbf{y}_1) & k(\mathbf{x}_2,\mathbf{y}_2) & \cdots & k(\mathbf{x}_2,\mathbf{y}_s) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_m,\mathbf{y}_1) & k(\mathbf{x}_m,\mathbf{y}_2) & \cdots & k(\mathbf{x}_m,\mathbf{y}_s) \end{bmatrix}.$$

Poniżej podajemy kilka przykładów funkcji kowariancji

- *Gaussian kernel* dla normy $\|\cdot\|$ i hiper-parametru l

$$k(\mathbf{x},\mathbf{y}) = \exp\left\{-\frac{1}{2l^2}\|\mathbf{x}-\mathbf{y}\|^2\right\}$$

- *Periodic kernel* dla normy $\|\cdot\|$ i hiper-parametrów l, p

$$k(\mathbf{x},\mathbf{y}) = \exp\left\{-\frac{2}{l^2}\sin^2\left(\frac{\pi}{p}\|\mathbf{x}-\mathbf{y}\|\right)\right\}$$

- *White noise kernel* dla hiper-parametru σ

$$k(\mathbf{x},\mathbf{y}) = \sigma^2\delta_{\mathbf{x},\mathbf{y}}$$

- *Matérn kernel* dla normy $\|\cdot\|$ i hiper-parametrów l, ν

$$k(\mathbf{x},\mathbf{y}) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}}{l}\|\mathbf{x}-\mathbf{y}\|\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{l}\|\mathbf{x}-\mathbf{y}\|\right),$$

  gdzie $\Gamma(x)$ to funkcja gamma Eulera, a $K_\nu(x)$ to zmodyfikowana funkcja Bessela 2-go rodzaju rzędu ν.

Dodatkowo suma lub iloczyn dwóch funkcji kowariancji oraz złożenie funkcji kowariancji z wielomianem o nieujemnych współczynnikach jest również funkcją kowariancji.

Procesem Gaussowskim (z ang. *Gaussian Process*) nazywamy rodzinę skalarnych zmiennych losowych indeksowanych przez punkty $\mathbf{x} \in \mathbb{R}^n$

$$\mathcal{GP} = \{f_{\mathbf{x}} \mid \mathbf{x} \in \mathbb{R}^n\}$$

taką że każdy skończony podzbiór $\mathcal{GP}$ ma łącznie wielowymiarowy rozkład normalny tj. dla dowolnego zbioru $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathbb{R}^n$ zachodzi

$$\begin{bmatrix} f_{\mathbf{x}_1} \\ \vdots \\ f_{\mathbf{x}_m} \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X).$$

Zauważmy, iż process Gaussowski możemy jednoznacznie zdefiniować podając przepisy na parametry $\boldsymbol{\mu}_X$ i $\boldsymbol{\Sigma}_X$ dla dowolnego zbioru X. W praktyce często przyjmujemy $\boldsymbol{\mu}_X = \mathbf{0}$, natomiast przepisem na macierz kowariancji może być zdefiniowana wyżej funkcja kowariancji $k(X, X)$ tj.

$$\begin{bmatrix} f_{\mathbf{x}_1} \\ \vdots \\ f_{\mathbf{x}_m} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, k(X, X)) \,.$$

Process Gaussowski daje nam w praktyce rozkład prawdopodobieństwa nad funkcjami $f : \mathbb{R}^n \mapsto \mathbb{R}$, których charakter jest określony przez jądro k (np. funkcja gładka dla jądra Gaussowskiego, okresowa dla jądra periodycznego, itp.). Zauważmy, że nie wnioskujemy tu o parametrach konkretnej rodziny funkcji (jak w przypadku regresji liniowej); interesuje nas jedynie rozkład predykcyjny. Załóżmy, iż w zadanych (lub dokładnie znanych) przez nas punktach $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ zaobserwowaliśmy wartości pewnej funkcji, o których zakładamy, iż pochodzą z procesu Gaussowskiego zadanego jądrem k, które wyraża nasze założenia a priori co do charakteru badanej funkcji

$$\mathbf{f}_X = \begin{bmatrix} f_{\mathbf{x}_1} \\ \vdots \\ f_{\mathbf{x}_m} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, k(X, X)) \,.$$

Powiedzmy, iż chcemy znać wartości $\mathbf{f}_Y$ tej funkcji w zadanych punktach $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_s\}$. Ponieważ założyliśmy, iż wartości funkcji pochodzą z procesu Gaussowskiego, więc rozkład łączny $\mathbf{f}_X$ i $\mathbf{f}_Y$ jest rozkładem normalnym

$$\begin{bmatrix} \mathbf{f}_X \\ \mathbf{f}_Y \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k(X, X) & k(X, Y) \\ k(Y, X) & k(Y, Y) \end{bmatrix}\right) .$$

Zauważmy, iż jest to instancja modelu Gaussowskiego, więc rozkład warunkowy $\mathbf{f}_Y \mid \mathbf{f}_X$ jest również rozkładem normalnym o parametrach

$$\begin{aligned} \boldsymbol{\mu} &= k(Y, X)k^{-1}(X, X)\mathbf{f}_X \\ \boldsymbol{\Sigma} &= k(Y, Y) - k(Y, X)k^{-1}(X, X)k(X, Y) \end{aligned} \quad .$$

Dodatkową niepewność związaną z pomiarem wartości $\mathbf{f}_X$ możemy uchwycić zmieniając postać jądra

$$k(\mathbf{x}, \mathbf{y}) \leftarrow k(\mathbf{x}, \mathbf{y}) + \mathcal{I}_X(\mathbf{x})\sigma^2\delta_{\mathbf{x},\mathbf{y}} \,,$$

gdzie σ jest hiper-parametrem określającym precyzję pomiaru. Oczywiście k jest dalej funkcją kowariancji, gdyż takie podstawienie powoduje jedynie dodanie dodatnich członów do pewnych elementów diagonalnych macierzy kowariancji, więc

macierz ta jest nadal symetryczna i dodatnio określona. Wówczas rozkład predykcyjny ma parametry

$$\begin{aligned}\boldsymbol{\mu} &= k(Y,X)\left[k(X,X)+\sigma^2\mathbf{1}\right]^{-1}\mathbf{f}_X \\ \boldsymbol{\Sigma} &= k(Y,Y)-k(Y,X)\left[k(X,X)+\sigma^2\mathbf{1}\right]^{-1}k(X,Y)\end{aligned}\,.$$

4.9 Wieloklasowa regresja logistyczna

Załóżmy, iż modelujemy obserwacje postaci $(t,\mathbf{x})$, gdzie $t \in \{\tau_1,\tau_2,\ldots,\tau_s\}$ to etykieta określająca przynależność do jednej z s klas, a $\mathbf{x}\in\mathbb{R}^n$ jest znanym (lub zadanym) przez nas dokładnie wektorem cech obiektu dla których zaobserwowaną klasą jest t. Zakładamy ponadto, iż prawdopodobieństwo przynależności do klasy τ_j (jednej z s klas) dla wektora cech $\mathbf{x}$ ma postać tzw. funkcji softmax

$$\pi_j(\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\mathrm{e}^{\mathbf{w}_j^\top\mathbf{x}}\,,$$

gdzie $\mathbf{w}_j$ są estymowanymi przez nas parametrami. Ze względu na warunek unormowania musimy mieć

$$\sum_{j=1}^{s}\pi_j = 1\,,$$

skąd stała normalizacyjna $Z(\mathbf{x})$ ma postać

$$Z(\mathbf{x}) = \sum_{j=1}^{s}\mathrm{e}^{\mathbf{w}_j^\top\mathbf{x}}\,.$$

Rozkład zmiennej losowej t jest w takim razie dyskretnym rozkładem wielopunktowym (z ang. *categorical distribution*) postaci

$$t \mid \mathbf{w}_1,\ldots,\mathbf{w}_s \sim \mathrm{Cat}(\pi_1(\mathbf{x}),\ldots,\pi_s(\mathbf{x}))\,.$$

Zauważmy, iż prawdopodobieństwo wylosowania etykiety t dla parametrów $\mathbf{w}_j$ możemy zapisać jako

$$p(t \mid \mathbf{w}_1,\ldots,\mathbf{w}_s) = \prod_{j=1}^{s}\pi_j(\mathbf{x})^{\delta(t,\tau_j)}\,.$$

Powiedzmy, że mamy obserwacje $D=(t_1,\ldots,t_m)$ dla znanych (lub zadanych) przez nas dokładnie wektorów cech $(\mathbf{x}_1,\ldots,\mathbf{x}_m)$. Funkcja wiarygodności ma wówczas postać

$$p(D \mid \mathbf{w}_1,\ldots,\mathbf{w}_s) = \prod_{i=1}^{m}p(t_i \mid \mathbf{w}_1,\ldots,\mathbf{w}_s) = \prod_{i=1}^{m}\prod_{j=1}^{s}\pi_j(\mathbf{x}_i)^{\delta(t_i,\tau_j)}\,.$$

Jako prior dla parametrów $\mathbf{w}_j$ przyjmiemy rozkład normalny z pewnym hiper-parametrem γ

$$\forall j \in \{1, \ldots, s\} : \mathbf{w}_j \sim \mathcal{N}(\mathbf{0}, \gamma^{-1}\mathbf{1}) \,.$$

W przypadku regresji logistycznej ograniczymy się do znalezienia estymaty MAP parametrów $\mathbf{w}_j$ tak, aby w przyszłości do nowego wektora cech $\mathbf{x}$ przyporządkować klasę o największym prawdopodobieństwie $\pi_j(\mathbf{x})$. Znalezienie estymaty MAP sprowadza się do znalezienia minimum zregularyzowanej funkcji kosztu

$$\begin{aligned}
\ell^*(D \mid \mathbf{w}_1, \ldots, \mathbf{w}_s) &= -\log[p(D \mid \mathbf{w}_1, \ldots, \mathbf{w}_s)\pi(\mathbf{w}_1, \ldots, \mathbf{w}_s)] \\
&= -\log\left[\prod_{k=1}^{s} \mathrm{e}^{-\frac{\gamma}{2}\mathbf{w}_k^\top \mathbf{w}_k} \prod_{i=1}^{m}\prod_{j=1}^{s} \pi_j(\mathbf{x}_i)^{\delta(t_i,\tau_j)}\right] \\
&= \frac{\gamma}{2}\sum_{j=1}^{s} \mathbf{w}_j^\top \mathbf{w}_j - \sum_{i=1}^{m}\sum_{j=1}^{s} \delta(t_i, \tau_j) \log \pi_j(\mathbf{x}_i) \,.
\end{aligned}$$

Niestety dla tak zdefiniowanej funkcji kosztu nie można znaleźć wzoru na minimum w postaci analitycznej, dlatego wykorzystamy numeryczny algorytm optymalizacji zwany spadkiem wzdłuż gradientu.

Algorytm spadku wzdłuż gradientu

1. Wybierz parametry początkowe $\mathbf{x}_1^{(0)}, \ldots, \mathbf{x}_m^{(0)}$
2. Powtarzaj

$$\begin{aligned}
\mathbf{x}_1^{(t+1)} &= \mathbf{x}_1^{(t)} - \epsilon_1 \frac{\partial f}{\partial \mathbf{x}_1}\bigg|_{\mathbf{x}_1^{(t)},\ldots,\mathbf{x}_m^{(t)}} \\
&\vdots \\
\mathbf{x}_m^{(t+1)} &= \mathbf{x}_m^{(t)} - \epsilon_m \frac{\partial f}{\partial \mathbf{x}_m}\bigg|_{\mathbf{x}_1^{(t)},\ldots,\mathbf{x}_m^{(t)}}
\end{aligned}$$

gdzie $\epsilon_1, \ldots, \epsilon_m$ to hiper-parametry zwane stałymi uczącymi (z ang. *learning rate*).

Zakładając $\epsilon_1 = \ldots = \epsilon_m = \epsilon$ i wprowadzając

$$\mathbf{X} := \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_m^\top \end{bmatrix}, \quad \frac{\partial f}{\partial \mathbf{X}} := \begin{bmatrix} \frac{\partial f}{\partial \mathbf{x}_1}^\top \\ \vdots \\ \frac{\partial f}{\partial \mathbf{x}_m}^\top \end{bmatrix}$$

możemy zapisać powyższe równania w kompaktowej formie

$$\mathbf{X}^{(t+1)} = \mathbf{X}^{(t)} - \epsilon \frac{\partial f}{\partial \mathbf{X}}\bigg|_{\mathbf{X}^{(t)}} .$$

Aby zminimalizować numerycznie funkcję kosztu ℓ^* stosując metodę spadku wzdłuż gradientu musimy obliczyć pochodne funkcji kosztu po parametrach $\mathbf{w}_j$

$$\frac{\partial \ell^*}{\partial \mathbf{w}_k} = \gamma \mathbf{w}_k - \sum_{i=1}^{m} \sum_{j=1}^{s} \delta(t_i, \tau_j) \frac{\partial}{\partial \mathbf{w}_k} \log \pi_j(\mathbf{x}_i) \,,$$

ale

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}_k} \log \pi_j(\mathbf{x}_i) &= \frac{1}{\pi_j(\mathbf{x}_i)} \frac{Z(\mathbf{x}_i) \frac{\partial \mathrm{e}^{\mathbf{x}_i^\top \mathbf{w}_j}}{\partial \mathbf{w}_k} - \mathrm{e}^{\mathbf{x}_i^\top \mathbf{w}_j} \frac{\partial Z(\mathbf{x}_i)}{\partial \mathbf{w}_k}}{Z^2(\mathbf{x}_i)} \\
&= \frac{Z(\mathbf{x}_i)}{\mathrm{e}^{\mathbf{x}_i^\top \mathbf{w}_j}} \frac{Z(\mathbf{x}_i) \mathbf{x}_i \mathrm{e}^{\mathbf{x}_i^\top \mathbf{w}_k} \delta_{jk} - \mathrm{e}^{\mathbf{x}_i^\top \mathbf{w}_j} \mathrm{e}^{\mathbf{x}_i^\top \mathbf{w}_k} \mathbf{x}_i}{Z^2(\mathbf{x}_i)} \\
&= \mathbf{x}_i \delta_{jk} - \mathbf{x}_i \pi_k(\mathbf{x}_i)
\end{aligned}$$

zatem

$$\frac{\partial \ell^*}{\partial \mathbf{w}_k} = \gamma \mathbf{w}_k - \sum_{i=1}^{m} \mathbf{x}_i \sum_{j=1}^{s} \delta(t_i, \tau_j) \delta_{jk} + \sum_{i=1}^{m} \mathbf{x}_i \pi_k(\mathbf{x}_i) \sum_{j=1}^{s} \delta(t_i, \tau_j) \,.$$

Zauważmy jednak, iż

$$\sum_{j=1}^{s} \delta(t_i, \tau_j) = 1 \,, \quad \sum_{j=1}^{s} \delta(t_i, \tau_j) \delta_{jk} = \delta(t_i, \tau_k) \,,$$

zatem ostatecznie

$$\frac{\partial \ell^*}{\partial \mathbf{w}_k} = \gamma \mathbf{w}_k + \sum_{i=1}^{m} \mathbf{x}_i \left[\pi_k(\mathbf{x}_i) - \delta(t_i, \tau_k) \right] .$$

Wprowadzając macierze

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_m^\top \end{bmatrix} , \quad \mathbf{W} = \begin{bmatrix} \mathbf{w}_1^\top \\ \vdots \\ \mathbf{w}_s^\top \end{bmatrix} ,$$

$$\mathbf{S} = \begin{bmatrix} \pi_1(\mathbf{x}_1) & \cdots & \pi_s(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \pi_1(\mathbf{x}_m) & \cdots & \pi_s(\mathbf{x}_m) \end{bmatrix} , \quad \mathbf{T} = \begin{bmatrix} \delta(t_1, \tau_1) & \cdots & \delta(t_1, \tau_s) \\ \vdots & \ddots & \vdots \\ \delta(t_m, \tau_1) & \cdots & \delta(t_m, \tau_s) \end{bmatrix}$$

możemy w takim razie zapisać zdefiniowaną wyżej macierz pochodnych wymaganych do algorytmu spadku wzdłuż gradient w kompaktowej formie jako

$$\frac{\partial \ell^*}{\partial \mathbf{W}} = (\mathbf{S} - \mathbf{T})^\top \mathbf{X}\,.$$

Zauważmy, iż zregularyzowana funkcja kosztu rośnie wraz ze wzrostem liczby obserwacji m. Wynika z tego, iż stała ucząca musi być zależna od liczby przykładów. Możemy na przykład stwierdzić, iż $\epsilon \leftarrow m^{-1}\epsilon$ i wówczas minimalizujemy tak naprawdę średni koszt ℓ^*/m.

4.10 Wnioskowanie metodami Monte Carlo

Całe wnioskowanie Bayesowskie opiera się na wyznaczaniu rozkładów a posteriori, które wyrażają naszą wiedzę o estymowanym parametrze. Do tej pory rozważaliśmy modele Bayesowskie dla których prior i wiarygodność były dane przez rozkłady normalne. Dzięki temu mogliśmy wyprowadzić analityczne wzory na parametry rozkładu a posteriori, który również był rozkładem normalnym. Dla wielu interesujących modeli nie jesteśmy jednak w stanie tego zrobić (np. w zagadnieniu regresji logistycznej ograniczyliśmy się jedynie do estymaty punktowej), gdyż obliczenie stałej normalizującej dla rozkładu $p(\theta \mid D)$ może wymagać obliczenia całki, której nie jesteśmy w stanie wyrazić w sposób jawny lub sumy po wykładniczo wielu elementach. Wnioskowanie Bayesowskie można jednak prowadzić w modelach, w których nie dysponujemy jawnym wzorem na gęstość prawdopodobieństwa rozkładu a posteriori. Okazuje się, iż do generowania próbek z rozkładu $p(\theta \mid D)$ wystarcza znajomość tego rozkładu z dokładnością do stałej normalizującej, a zatem wystarczy znać rozkład łączny $p(\theta, D) = p(D \mid \theta)\pi(\theta)$. Generowanie próbek z kolei wystarcza natomiast, na mocy silnego prawa wielkich liczb, do szacowania wartości średnich dowolnych funkcji estymowanego parametru θ. Przypomnijmy, iż na mocy silnego prawa wielkich liczb ciąg średnich częściowych $(\overline{X}_n)$ ciągu zmiennych losowych (X_n) i.i.d. z rozkładu $X \sim \mathcal{D}$ jest zbieżny z prawdopodobieństwem 1 do wartości oczekiwanej $\mathbb{E}[X]$ tj.

$$P\left(\lim_{n\to\infty} \overline{X}_n = \mathbb{E}[X]\right) = 1\,.$$

Wartość oczekiwaną $\mathbb{E}[X]$ możemy zatem przybliżyć średnią $\overline{X}_n$ z dużej ilości próbek.

Wnioskowanie Monte Carlo pozwala nam szacować różne wielkości w tzw. hierarchicznych modelach Bayesowskich (z ang. *Bayesian hierarchical modeling*). Rozważmy jeszcze raz przykład regresji liniowej w ujęciu Bayesowskim, ale rozważmy

teraz model postaci

$$\begin{aligned} \sigma^2 &\sim \mathcal{D}(\lambda) \\ \mathbf{w} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ y \mid \mathbf{w}, \sigma^2 &\sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2) \end{aligned} ,$$

gdzie $\lambda, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$ są pewnymi hiper-parametrami. Dla takiego modelu nie możemy w ogólności znaleźć jawnej postaci rozkładu a posteriori. Jeśli jednak umiemy generować próbki z rozkładu łącznego

$$Z \cdot p(\mathbf{w}, \sigma^2 \mid D) = p(D, \mathbf{w}, \sigma^2) = p(D \mid \mathbf{w}, \sigma^2)\pi(\mathbf{w})\pi(\sigma^2)$$

to wszystkie interesujące wielkości możemy oszacować jako odpowiednie średnie. Pozostaje pytanie w jaki sposób generować próbki ze skomplikowanych rozkładów prawdopodobieństwa, których gęstości znamy jedynie z dokładnością do stałej normalizującej. Poniżej przedstawimy dwa algorytmy próbkowania: algorytm IS oraz Metropolisa–Hastingsa będący szczególną realizacją całej rodziny algorytmów próbkowania zwanych Markov Chain Monte Carlo (MCMC).

4.10.1 Algorytm Importance Sampling (IS)

Załóżmy, iż chcemy obliczyć wartość oczekiwaną pewnej funkcji zmiennej losowej $\mathbf{x}$ względem skomplikowanego rozkładu prawdopodobieństwa $p(\mathbf{x})$, który znamy jedynie z dokładnością do stałej normalizującej

$$p(\mathbf{x}) = \frac{1}{Z_p}\tilde{p}(\mathbf{x})$$

tj. szukamy

$$\mathbb{E}_p[f(\mathbf{x})] = \int f(\mathbf{x})p(\mathbf{x})\mathrm{d}^n\mathbf{x}\,.$$

Jeśli umiemy generować próbki $\mathbf{x}$ z innego (prostszego) rozkładu $q(\mathbf{x})$, który nazywamy rozkładem proponującym kandydatów (z ang. *proposal distribution*) to możemy zapisać

$$\begin{aligned} \mathbb{E}_p[f(\mathbf{x})] &= \int\limits_{\mathbb{R}^n} f(\mathbf{x})p(\mathbf{x})\mathrm{d}^n\mathbf{x} = \int\limits_{\mathbb{R}^n} f(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}q(\mathbf{x})\mathrm{d}^n\mathbf{x} \\ &= \mathbb{E}_q\left[f(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}\right] = \frac{Z_q}{Z_p}\mathbb{E}_q\left[f(\mathbf{x})\frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})}\right] . \end{aligned} .$$

Zakładamy tutaj, iż nośnik rozkładu p zawiera się w nośniku q tj. $\operatorname{supp} p \subseteq \operatorname{supp} q$. Stosunek stałych Z_p/Z_q również możemy oszacować z próbek z q, gdyż mamy

$$Z_p = \int\limits_{\mathbb{R}^n} \tilde{p}(\mathbf{x})\,\mathrm{d}^n\mathbf{x} = Z_q \int\limits_{\mathbb{R}^n} \frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})}q(\mathbf{x})\,\mathrm{d}^n\mathbf{x} = Z_q\mathbb{E}_q\left[\frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})}\right] ,$$

skąd ostatecznie

$$\mathbb{E}_p[f(\mathbf{x})] = \frac{\mathbb{E}_q\left[f(\mathbf{x})\frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})}\right]}{\mathbb{E}_q\left[\frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})}\right]}\,.$$

Jeśli z rozkładu q wygenerowaliśmy próbki $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ to na mocy silnego prawa wielkich liczb mamy

$$\mathbb{E}_p[f(\mathbf{x})] \approx \frac{\sum_{i=1}^m f(\mathbf{x}_i)\frac{\tilde{p}(\mathbf{x}_i)}{\tilde{q}(\mathbf{x}_i)}}{\sum_{i=1}^m \frac{\tilde{p}(\mathbf{x}_i)}{\tilde{q}(\mathbf{x}_i)}} = \sum_{i=1}^m \lambda_i f(\mathbf{x}_i)\,,$$

gdzie

$$\lambda_i = \frac{\tilde{p}(\mathbf{x}_i)/\tilde{q}(\mathbf{x}_i)}{\sum_{j=1}^m \tilde{p}(\mathbf{x}_j)/\tilde{q}(\mathbf{x}_j)}\,.$$

Algorytm Importance Sampling jest prostym algorytmem Monte Carlo, który ma jeden zasadniczy problem. W jaki sposób mamy wybrać rozkład proponujący kandydatów q? Pewną odpowiedź na to pytanie sugeruje analiza wariancji statystyki

$$\overline{f}_m(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \frac{1}{m}\sum_{i=1}^m \frac{f(\mathbf{x}_i)p(\mathbf{x}_i)}{q(\mathbf{x}_i)}$$

dla $\mathbf{x}_i \sim q$ i zakładając dla uproszczenia, iż f jest funkcją skalarną mamy

$$\mathrm{Var}[\overline{f}_m] = \frac{1}{m}\mathrm{Var}_q\left[f(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}\right] = \frac{1}{m}\int\limits_{\mathbb{R}^n} \frac{(f(\mathbf{x})p(\mathbf{x}) - \mu_f q(\mathbf{x}))^2}{q(\mathbf{x})}\,\mathrm{d}^n\mathbf{x}\,.$$

Chcemy oczywiście, aby wariancja była jak najmniejsza, gdyż wówczas mała liczba próbek da dobre przybliżenie wartości oczekiwanej. Rozkład proponujący kandydatów powinien być zatem proporcjonalny do $f(\mathbf{x})p(\mathbf{x})$, co może być trudne do praktycznego zrealizowania.

4.10.2 Algorytm Metropolisa–Hastingsa

Cała klasa algorytmów próbkowania MCMC opiera się na idei wyrażenia generowania próbek jako ewolucji pewnego łańcucha Markowa. Łańcuchem Markowa nazywamy ciąg zmiennych losowych (X_t) o wartościach w $\mathbb{R}^n$ taki, że spełnione jest kryterium Markowa

$$\forall A \subset \mathbb{R}^n : P(X_t \in A \mid X_{t-1} = \mathbf{x}_{t-1}, \ldots, X_0 = \mathbf{x}_0) = P(X_t \in A \mid X_{t-1} = \mathbf{x}_{t-1})\,.$$

Elementy ciągu nazywamy stanami łańcucha Markowa. Dany łańcuch jest zadany jednoznacznie przez podanie gęstości prawdopodobieństwa przejścia łańcucha ze

stanu $\mathbf{x} \to \mathbf{y}$, którą będziemy oznaczać przez $\pi(\mathbf{y} \mid \mathbf{x})$ (zakładamy, iż prawdopodobieństwo przejścia jest niezależne od chwili t – łańcuch taki nazywamy jednorodnym). Funkcja π spełnia oczywiście warunek unormowania

$$\int_{\mathbb{R}^n} \pi(\mathbf{y} \mid \mathbf{x}) \, \mathrm{d}^n\mathbf{y} \ ,$$

istotnie prawdopodobieństwo przejścia gdziekolwiek ze stanu $\mathbf{x}$ jest równe 1. Będziemy zakładać dodatkowo, iż $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n : \pi(\mathbf{y} \mid \mathbf{x}) > 0$. Rozkład $p(\mathbf{x})$ łańcucha Markowa (tj. rozkład prawdopodobieństwa z którego losujemy stan łańcucha w danej chwili t) z daną funkcją przejścia π nazwiemy rozkładem stacjonarnym tego łańcucha iff

$$p(\mathbf{y}) = \int_{\mathbb{R}^n} \pi(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}) \, \mathrm{d}^n\mathbf{x} \ .$$

Rozkład stacjonarny danego łańcucha oznaczymy przez $p^*(\mathbf{x})$. Zauważmy, iż jeśli stan początkowy łańcucha X_0 pochodzi z rozkładu stacjonarnego p^* to każdy kolejny stan X_t również pochodzi z rozkładu stacjonarnego. Jeśli z kolei stan początkowy pochodzi z jakiegoś innego rozkładu p_0 to rozkład łańcucha w chwili t jest dany przez relację rekurencyjną

$$p_t(\mathbf{y}) = \int_{\mathbb{R}^n} \pi(\mathbf{y} \mid \mathbf{x}) p_{t-1}(\mathbf{x}) \, \mathrm{d}^n\mathbf{x} \ , \quad \text{dla } t > 1.$$

Rozkładem granicznym łańcucha Markowa nazwiemy granicę w sensie zbieżności punktowej

$$\lim_{t \to \infty} p_t(\mathbf{x}) \, .$$

Przy podanych wyżej założeniach istnieje twierdzenie, które mówi iż taki łańcuch Markowa posiada jednoznaczny rozkład stacjonarny tożsamy z rozkładem granicznym. Ponadto warunkiem wystarczającym, aby dany rozkład $p(\mathbf{x})$ był rozkładem stacjonarnym łańcucha Markowa jest

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n : \pi(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}) = \pi(\mathbf{x} \mid \mathbf{y}) p(\mathbf{y}) \, ,$$

co wynika z scałkowania powyższego równania

$$\int_{\mathbb{R}^n} \pi(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}) \, \mathrm{d}^n\mathbf{x} = \int_{\mathbb{R}^n} \pi(\mathbf{x} \mid \mathbf{y}) p(\mathbf{y}) \, \mathrm{d}^n\mathbf{x} = p(\mathbf{y}) \int_{\mathbb{R}^n} \pi(\mathbf{x} \mid \mathbf{y}) \, \mathrm{d}^n\mathbf{x} = p(\mathbf{y}) \, .$$

Kryterium to nazywamy kryterium lokalnego balansu (z ang. *detailed balance condition*).

Podstawowa idea wykorzystania łańcuchów Markowa do generowania próbek ze skomplikowanego rozkładu p jest więc następująca: tworzymy łańcuch Markowa opisany powyżej, dla którego p jest rozkładem stacjonarnym, wówczas rozpoczynając w dowolnym dopuszczalnym stanie początkowym X_0 po wykonaniu dużej liczby kroków (etap ten nazywamy okresem przejściowym z ang. *burn-in period*) stan X_t (dla $t \gg 1$) tego łańcucha będzie w przybliżeniu pochodził z rozkładu granicznego p (nie jest jednak prosto stwierdzić po jak długim okresie przejściowym przybliżenie to jest wystarczająco dobre). Aby otrzymać z takiej procedury próbki prawdziwie i.i.d. każda z próbek musiałaby pochodzić z ponownego uruchomienia takiego łańcucha. Oczywiście jest to nieefektywne, więc w praktyce generujemy próbki z jednego łańcucha po prostu odrzucając pewne z nich tak aby uniknąć znaczących korelacji.

Pozostaje pytanie jak skonstruować funkcję przejścia $\pi(\mathbf{y} \mid \mathbf{x})$ dla danego rozkładu granicznego $p(\mathbf{x})$. Podstawową konstrukcję podaje algorytm Metropolisa–Hastingsa:

1. Jako stan początkowy przyjmij dowolną dopuszczalną wartość $\mathbf{x} \leftarrow \mathbf{x}_0$.

2. Powtarzaj:

   (a) Będąc w aktualnym stanie $\mathbf{x}$ z prostego rozkładu proponującego kandydatów $q(\mathbf{y} \mid \mathbf{x})$ wylosuj kandydata $\mathbf{y}$ na wartość łańcucha w kolejnym stanie.

   (b) Z prawdopodobieństwem

   $$r(\mathbf{y} \mid \mathbf{x}) = \min\left\{1, \frac{p(\mathbf{y})q(\mathbf{x} \mid \mathbf{y})}{p(\mathbf{x})q(\mathbf{y} \mid \mathbf{x})}\right\}$$

   zaakceptuj kandydata jako nowy stan i przejdź do stanu $\mathbf{y}$. W przeciwnym razie pozostać w stanie $\mathbf{x}$

Funkcja przejścia ma zatem postać

$$\pi_{\mathrm{MH}}(\mathbf{y} \mid \mathbf{x}) = q(\mathbf{y} \mid \mathbf{x}) r(\mathbf{y} \mid \mathbf{x})\,.$$

Pozostaje tylko wykazać, iż spełnione jest kryterium lokalnego balansu. Istotnie mamy

$$\begin{aligned} \pi_{\mathrm{MH}}(\mathbf{y} \mid \mathbf{x}) p(\mathbf{x}) &= \min\{q(\mathbf{y} \mid \mathbf{x})p(\mathbf{x}), q(\mathbf{x} \mid \mathbf{y})p(\mathbf{y})\} \\ \pi_{\mathrm{MH}}(\mathbf{x} \mid \mathbf{y}) p(\mathbf{y}) &= \min\{q(\mathbf{x} \mid \mathbf{y})p(\mathbf{y}), q(\mathbf{y} \mid \mathbf{x})p(\mathbf{x})\} \end{aligned}\,,$$

skąd $\pi_{\mathrm{MH}}(\mathbf{y} \mid \mathbf{x})p(\mathbf{x}) = \pi_{\mathrm{MH}}(\mathbf{x} \mid \mathbf{y})p(\mathbf{y})$. Zauważmy, iż nie musimy znać $p(\mathbf{x})$ z dokładnością do stałej normalizującej, gdyż

$$\frac{p(\mathbf{y})}{p(\mathbf{x})} = \frac{\tilde{p}(\mathbf{y})/Z_p}{\tilde{p}(\mathbf{x})/Z_p} = \frac{\tilde{p}(\mathbf{y})}{\tilde{p}(\mathbf{x})}\,.$$

Poza algorytmem Metropolisa–Hastingsa jest wiele innych algorytmów z rodziny MCMC. Większość z nich implementuje konkretny sposób generowania (zostawiając resztę struktury) tak, aby zmniejszyć korelację po okresie przejściowym i przyspieszyć zbieżność. Standardowo wykorzystywanymi algorytmami z tej klasy są algorytmy HMC (*Hamiltonian Monte Carlo*) oraz NUTS (*No U-Turn Sampler*).

5 Sieci neuronowe

Podstawowym elementem każdej sieci neuronowej jest pojedynczy neuron, który możemy traktować jako odwzorowanie postaci $z : \mathbb{R}^n \mapsto \mathbb{R}$ będące złożeniem pewnego odwzorowania nieliniowego $f : \mathbb{R} \mapsto \mathbb{R}$ z odwzorowaniem afinicznym $a : \mathbb{R}^n \mapsto \mathbb{R}$, $a(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ tj.

$$z(\mathbf{x}) = (f \circ a)(\mathbf{x}) = f\left(\mathbf{w}^\top \mathbf{x} + b\right) .$$

W praktyce wszystkie neurony sieci używają tej samej nieliniowej funkcji f zwanej funkcją aktywacji (z ang. *activation function*) i najczęściej są to funkcje ReLU, GELU lub funkcje sigmoidalne:

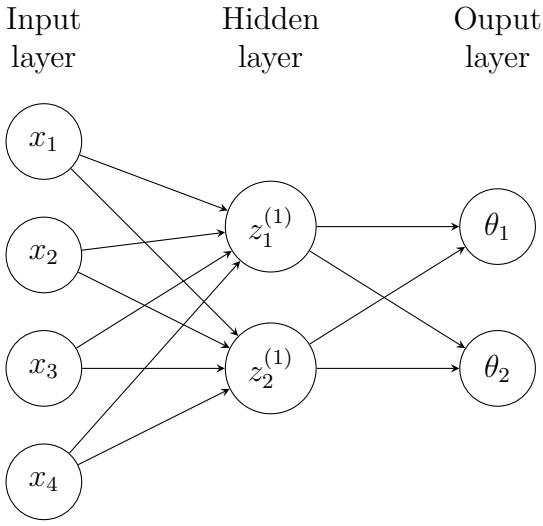
$$\begin{aligned} \mathrm{ReLU}(x) &:= \max(0, x) , \\ \mathrm{GELU}(x) &:= x\Phi(x) , \end{aligned}$$

gdzie $\Phi(x)$ to dystrybuanta standardowego rozkładu normalnego. Pojedyncze neurony są następnie łączone w sieci w określony sposób tworząc daną architekturę sieci neuronowej.

5.1 Architektura MLP

Opis sieci neuronowych zaczniemy od architektury MLP (z ang. *Multilayer Perceptron*). Sieć MLP składa się z równoległych warstw neuronów, przy czym połączenia występują tylko między neuronami w sąsiednich warstwach i nie ma połączeń między neuronami w obrębie jednej warstwy. Pierwszą warstwę sieci nazywamy warstwą wejściową (z ang. *input layer*), ostatnią – warstwą wyjściową (z ang. *output layer*), a pozostałe nazywamy warstwami ukrytymi (z ang. *hidden layers*).

Zauważmy, iż opisane wcześniej modele regresji liniowej i wieloklasowej regresji logistycznej są przykładamy najprostszych sieci MLP bez żadnych warstw ukrytych. Ich graficzne reprezentacje jako sieci MLP zamieszczono na Rysunku 1a i 1b. Zauważmy, iż wyjściem sieci są parametry docelowego rozkładu prawdopodobieństwa nad obserwacjami tj. odpowiednio wartość oczekiwana μ w przypadku regresji liniowej i prawdopodobieństwa π_i każdej z klas rozkładu kategorialnego w przypadku regresji logistycznej. W przypadku regresji liniowej funkcja aktywacji



neuronu w warstwie wyjściowej to po prostu funkcja identycznosciowa, natomiast w przypadku regresji logistycznej jest to funkcja soft-max.

Przejdźmy teraz do matematycznego opisu architektury MLP. Dla danej funkcji $f : \mathbb{R} \mapsto \mathbb{R}$ przez zapis $\mathbf{f}(\mathbf{x})$ dla $\mathbf{x} \in \mathbb{R}^n$ będziemy rozumieli macierz kolumnową

$$\mathbf{f}(\mathbf{x}) = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}.$$

Dodatkowo zdefiniujmy dodatkowo odwzorowanie $\mathbf{a} : \mathbb{R}^n \mapsto \mathbb{R}^m$ jako

$$\mathbf{a}(\mathbf{x}) = \begin{bmatrix} \mathbf{w}_1^\top \\ \vdots \\ \mathbf{w}_m^\top \end{bmatrix} \mathbf{x} + \mathbf{b} = \mathbf{W}\mathbf{x} + \mathbf{b}\,,$$

gdzie $\mathbf{W} \in \mathbb{M}_{m\times n}(\mathbb{R}), \mathbf{b} \in \mathbb{R}^m$.

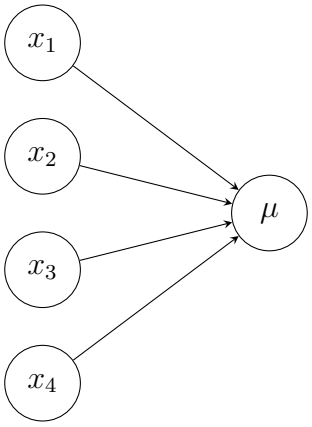
Oznaczmy przez $n_0, n_1, \ldots, n_{s-1}, n_s$ liczby neuronów w kolejnych warstwach, natomiast przez g funkcję aktywacji warstwy wyjściowej. Wyjście sieci MLP jest zatem opisane przez następujące złożenie funkcji

$$\boldsymbol{\theta}(\mathbf{x}; \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s) = (\mathbf{g} \circ \mathbf{a}_s \circ \mathbf{f}_{s-1} \circ \mathbf{a}_{s-1} \circ \ldots \circ \mathbf{f}_1 \circ \mathbf{a}_1)(\mathbf{x})\,,$$

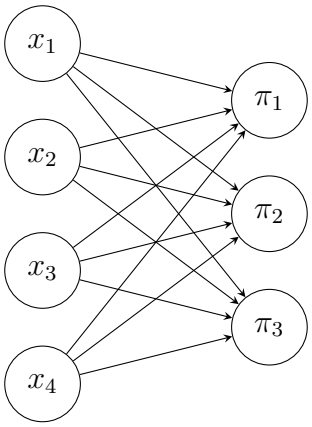
gdzie

$$\mathbf{a}_k(\mathbf{x}) = \mathbf{W}_k\mathbf{x} + \mathbf{b}_k\,,$$

przy czym $\mathbf{W}_k \in \mathbb{M}_{n_k\times n_{k-1}}(\mathbb{R})$ oraz $\mathbf{f}_k : \mathbb{R}^{n_k} \mapsto \mathbb{R}^{n_k}$. Chcemy zatem wnioskować o parametrach $\mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s$ zakładając, iż rozkład warunkowy nad



(a) Graficzna reprezentacja regresji liniowej jako najprostszej sieci MLP



(b) Graficzna reprezentacja wieloklasowej regresji logistycznej jako najprostszej sieci MLP

obserwacjami dla wektora zmiennych objaśniających $\mathbf{x}$ ma postać

$$y \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s \sim \mathcal{D}(\boldsymbol{\theta}(\mathbf{x}))\,.$$

dla pewnej rodziny rozkładów prawdopodobieństwa $\mathcal{D}$.

5.1.1 Wsteczna propagacja błędu

Zajmiemy się najpierw problemem znalezienia estymaty punktowej MLE dla parametrów sieci MLP. Załóżmy, iż mamy dane obserwacje iid $D = \{(y_1, \mathbf{x}_1), \ldots, (y_m, \mathbf{x}_m)\}$. Wiarygodność ma postać

$$p(D \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s) = \prod_{i=1}^{m} p(y_i \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s)$$

skąd funkcja kosztu (zanegowana logarytmiczna funkcja wiarygodności)

$$\ell(D \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s) = -\sum_{i=1}^{m} \log p(y_i \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s)\,.$$

Zauważmy więc, iż dla dowolnego modelu statystycznego funkcja ta ma postać sumy po wszystkich przykładach w zbiorze uczącym D

$$\ell(D \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s) = \sum_{i=1}^{m} \ell(y_i \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s)\,.$$

Do znalezienia estymaty MLE parametrów sieci neuronowej musimy zminimalizować powyższą funkcję, a zatem potrzebny nam jest algorytm efektywnego obliczania pochodnych $\ell(y_i \mid \ldots)$ po parametrach $\mathbf{W}_k, \mathbf{b}_k$. Zauważmy, że zachodzi

$$\frac{\partial \ell}{\partial \mathbf{W}_k} = \frac{\partial \ell}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{a}_s} \frac{\partial \mathbf{a}_s}{\partial \mathbf{f}_{s-1}} \frac{\partial \mathbf{f}_{s-1}}{\partial \mathbf{a}_{s-1}} \cdots \frac{\partial \mathbf{a}_k}{\partial \mathbf{W}_k}$$

jednakże

$$\frac{\partial \mathbf{f}_k}{\partial \mathbf{a}_k} = \begin{bmatrix} f'([\mathbf{a}_k]_1) & 0 & \ldots & 0 \\ 0 & f'([\mathbf{a}_k]_2) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & f'([\mathbf{a}_k]_{n_k}) \end{bmatrix} = \operatorname{diag}(\mathbf{f}'_k(\mathbf{a}_k))$$

oraz

$$\frac{\partial \mathbf{a}_k}{\partial \mathbf{f}_{k-1}} = \mathbf{W}_k\,, \quad \frac{\partial \mathbf{a}_k}{\partial \mathbf{W}_k} = \begin{bmatrix} \mathbf{f}_{k-1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{f}_{k-1} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{f}_{k-1} \end{bmatrix}$$

zatem

$$\frac{\partial \ell}{\partial \mathbf{W}_k} = \mathbf{f}_{k-1} \underbrace{\left[\frac{\partial \ell}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{a}_s} \cdot \mathbf{W}_s \cdot \operatorname{diag}(\mathbf{f}'_{s-1}(\mathbf{a}_{s-1})) \cdots \mathbf{W}_{k+1} \cdot \operatorname{diag}(\mathbf{f}'_k(\mathbf{a}_k)) \right]}_{\boldsymbol{\delta}_k}\,.$$

Zauważmy, że możemy obliczać $\boldsymbol{\delta}_k$ rekurencyjnie jako

$$\boldsymbol{\delta}_{k-1} = \boldsymbol{\delta}_k \cdot \mathbf{W}_k \cdot \operatorname{diag}(\mathbf{f}'_{k-1}(\mathbf{a}_{k-1}))\,, \quad \boldsymbol{\delta}_s = \frac{\partial \ell}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{a}_s}\,.$$

W przypadku wyrazu wolnego (*bias*) mamy natomiast analogicznie

$$\frac{\partial \ell}{\partial \mathbf{b}_k} = \frac{\partial \ell}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{a}_s} \frac{\partial \mathbf{a}_s}{\partial \mathbf{f}_{s-1}} \frac{\partial \mathbf{f}_{s-1}}{\partial \mathbf{a}_{s-1}} \cdots \frac{\partial \mathbf{a}_k}{\partial \mathbf{b}_k}$$

ponieważ jednak

$$\frac{\partial \mathbf{a}_k}{\partial \mathbf{b}_k} = \mathbf{I}\,,$$

więc

$$\frac{\partial \ell}{\partial \mathbf{b}_k} = \boldsymbol{\delta}_k\,.$$

Możemy zatem zapisać algorytm obliczania pochodnych funkcji kosztu po parametrach sieci neuronowej zwany algorytmem wstecznej propagacji błędu (z ang. *error backpropagation*)

Algorytm wstecznej propagacji błędu

1. Dla przykładu $(y_i, \mathbf{x}_i)$ dokonaj propagacji naprzód sieci MLP i zapamiętaj wartości funkcji $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_s$ i funkcji aktywacji $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{s-1}, \mathbf{g}$.

2. Wyznacz rekurencyjnie i zapamiętaj wartości $\boldsymbol{\delta}_k$ korzystając ze wstecznej propagacji

$$\boldsymbol{\delta}_{k-1} = \boldsymbol{\delta}_k \cdot \mathbf{W}_k \cdot \operatorname{diag}(\mathbf{f}'_{k-1}(\mathbf{a}_{k-1}))\,, \quad \boldsymbol{\delta}_s = \frac{\partial \ell}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{a}_s}\,.$$

3. Wyznacz odpowiednie pochodne korzystając z

$$\frac{\partial \ell}{\partial \mathbf{W}_k} = \mathbf{f}_{k-1}\boldsymbol{\delta}_k\,, \quad \frac{\partial \ell}{\partial \mathbf{b}_k} = \boldsymbol{\delta}_k\,.$$

W powyższym wzorze $\mathbf{f}_0 = \mathbf{x}_i$.

Powyższy algorytm wyznacza pochodną funkcji kosztu dla pojedynczego przykładu. Jeśli używamy serii przykładów D (tzw. *batch*) to oczywiście zachodzi

$$\frac{\partial \ell(D \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s)}{\partial \mathbf{W}_k} = \sum_{i=1}^{m} \frac{\partial \ell(y_i \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s)}{\partial \mathbf{W}_k}\,,$$

$$\frac{\partial \ell(D \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s)}{\partial \mathbf{b}_k} = \sum_{i=1}^{m} \frac{\partial \ell(y_i \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s)}{\partial \mathbf{b}_k}\,,$$

więc powyższy algorytm wykonujemy dla każdego przykładu i dodajemy wyniki. Problemy regresji liniowej i logistycznej na sieci MLP różnią się jedynie funkcją aktywacji g warstwy wyjściowej i używaną funkcją kosztu. W przypadku regresji liniowej mamy

$$g(x) = x\,, \quad \ell(y_i \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s) = \frac{1}{2}(y_i - g)^2\,,$$

skąd

$$\delta_s = g - y_i \,.$$

Natomiast w przypadku regresji logistycznej mamy

$$\mathbf{g}(\mathbf{x}) = \frac{1}{\sum_{i=1}^{n} \mathrm{e}^{x_i}} \begin{bmatrix} \mathrm{e}^{x_1} \\ \vdots \\ \mathrm{e}^{x_n} \end{bmatrix},$$

$$\ell(y_i \mid \mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s) = -\sum_{j=1}^{c} \delta(y_i, \tau_j) \log[\mathbf{g}]_j \,,$$

skąd

$$\boldsymbol{\delta}_s = \begin{bmatrix} \frac{\delta(y_i,\tau_1)}{[\mathbf{g}]_1} & \cdots & \frac{\delta(y_i,\tau_c)}{[\mathbf{g}]_c} \end{bmatrix} \left[[\mathbf{g}]_i [\mathbf{g}]_j - \delta_{ij} [\mathbf{g}]_j \right]_{c \times c} = \mathbf{g}^\top - \left[\delta(y_i, \tau_1) \ldots \delta(y_i, \tau_c) \right].$$

Mając algorytm efektywnego obliczania pochodnych funkcji kosztu, funkcję minimalizujemy korzystając z algorytmu spadku wzdłuż gradientu. W przypadku sieci neuronowych funkcja kosztu nie jest funkcją ściśle wypukłą, więc algorytm spadku wzdłuż gradientu nie znajdzie globalnego minimum; możemy liczyć jedynie na znalezienie minimum lokalnego. Istnieją trzy podstawowe algorytmy spadku wzdłuż gradientu: seryjny spadek wzdłuż gradientu (z ang. *Batch Gradient Descent (BGD)*), stochastyczny spadek wzdłuż gradientu (z ang. *Stochastic Gradient Descent (SGD)*) oraz mini-seryjny spadek wzdłuż gradientu (z ang. *Mini-Batch Gradient Descent (mBGD)*)

Algorytm BGD

1. Wybierz początkowe wartości parametrów $\mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s$.
2. Powtarzaj przez N epok:
   (a) Oblicz pochodne funkcji kosztu ℓ zsumowane po wszystkich przykładach z batcha D korzystając z algorytmu wstecznej propagacji błędu.
   (b) Zaktualizuj wartości parametrów zgodnie z:

$$\mathbf{W}_k = \mathbf{W}_k - \frac{\epsilon}{|D|} \frac{\partial \ell}{\partial \mathbf{W}_k}$$

$$\mathbf{b}_k = \mathbf{b}_k - \frac{\epsilon}{|D|} \frac{\partial \ell}{\partial \mathbf{b}_k}$$

Algorytm SGD

1. Wybierz początkowe wartości parametrów $\mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s$.
2. Powtarzaj przez N epok:
   Dla każdego przykładu $(y_i, \mathbf{x}_i) \in D$:
   (a) Oblicz pochodne funkcji kosztu ℓ dla przykładu $(y_i, \mathbf{x}_i)$ korzystając z algorytmu wstecznej propagacji błędu.
   (b) Zaktualizuj wartości parametrów zgodnie z:

$$\mathbf{W}_k = \mathbf{W}_k - \frac{\epsilon}{|D|}\frac{\partial \ell}{\partial \mathbf{W}_k}$$
$$\mathbf{b}_k = \mathbf{b}_k - \frac{\epsilon}{|D|}\frac{\partial \ell}{\partial \mathbf{b}_k}$$

Algorytm mBGD

1. Wybierz początkowe wartości parametrów $\mathbf{W}_1, \ldots, \mathbf{W}_s, \mathbf{b}_1, \ldots, \mathbf{b}_s$.
2. Podziel dane treningowe w D na K rozłącznych mini-batchy $D_1, \ldots, D_K$ jednakowej wielkości.
3. Powtarzaj przez N epok:
   Dla każdego mini-batcha D_i:
   (a) Oblicz pochodne funkcji kosztu ℓ zsumowane po wszystkich przykładach z mini-batcha D_i korzystając z algorytmu wstecznej propagacji błędu.
   (b) Zaktualizuj wartości parametrów zgodnie z:

$$\mathbf{W}_k = \mathbf{W}_k - \frac{\epsilon}{|D|}\frac{\partial \ell}{\partial \mathbf{W}_k}$$
$$\mathbf{b}_k = \mathbf{b}_k - \frac{\epsilon}{|D|}\frac{\partial \ell}{\partial \mathbf{b}_k}$$

5.1.2 Regularyzacja w sieciach neuronowych

1. **Consistent Gaussian Priors**
   Analogicznie jak w przypadku prostych modeli liniowych jedną z możliwości regularyzacji jest dodanie do funkcji kosztu ℓ czynnika regularyzującego

zawierającego kwadraty składowych wektorów wag postaci $\mathrm{tr}\big(\mathbf{W}_k^\top \mathbf{W}_k\big)$ przy czym w ogólności przyjmujemy iż mamy s różnych współczynników λ_k określających siły regularyzacji dla wag łączących poszczególne warstwy. Zregularyzowana funkcja kosztu ma więc w ogólności postać

$$\ell^* = \ell + \frac{1}{2}\sum_{k=1}^{s} \lambda_k \,\mathrm{tr}\big(\mathbf{W}_k^\top \mathbf{W}_k\big)\,.$$

Dzięki takiej postaci funkcja kosztu zachowuje własność niezmienniczości względem skalowania. Istotnie zauważmy, iż w przypadku niezregularyzowanej funkcji kosztu jeśli wektor w warstwie wejściowej pomnożymy przez pewien skalar α to wyjście sieci pozostanie niezmienione jeśli wagi w pierwszej warstwie ukrytej pomnożymy przez α^{-1}. Analogicznie wyjście nie zmieni się jeśli wektor w warstwie wyjściowej pomnożymy przez β, a wagi łączące dwie ostatnie warstwy pomnożymy przez β^{-1}. W szczególności możemy wykonać obie te operacje dla jednej sieci i również nie zmienimy wyjścia. Zauważmy jednak, iż gdybyśmy na sztywno założyli, iż współczynniki λ_1 i λ_s są takie same to zregularyzowana funkcja nie byłaby niezmiennicza względem takiego skalowania. Dzięki różnym współczynnikom pozostaje niezmiennicza jeśli pomnożymy współczynniki λ_1, λ_s odpowiednio przez $\sqrt{\alpha}$ i $\sqrt{\beta}$.

2. **Early stopping**
   Innym podejściem do regularyzacji sieci neuronowych jest procedura *early stopping*. Polega ona na zatrzymaniu procesu uczenia (optymalizacji funkcji kosztu) w momencie, w którym wartość funkcji kosztu na wydzielonym ze zbioru treningowego zbiorze walidacyjnym zaczyna rosnąć. Unikamy w ten sposób przeuczenia modelu.

3. **Niezmienniki**
   W wielu zastosowaniach uczenia maszynowego wiemy, że predykcje modelu powinny być niezmienione jeśli dane wejściowe poddamy pewnej transformacji np. w problemie rozpoznawania ręcznie pisanych cyfr odpowiedź nie powinna zależeć od miejsca na obrazku, w którym znajduje się cyfra (niezmienniczość translacyjna) oraz od skali, w której została napisana (niezmienniczość skalowania). Istnieje kilka możliwych rozwiązań pozwalających uwzględnić te cechy

   (a) Jeśli znamy przekształcenia względem których predykcje powinny pozostawać niezmienione możemy rozszerzyć nasz zbiór treningowy o przykłady z początkowego zbioru poddane tym przekształceniom.

   (b) Możemy dokonać preprocessingu danych w taki sposób aby wyekstrahować te cechy, które są niezmiennicze względem danych przekształceń (*feature extraction*).

(c) Możemy w końcu wbudować tę niezmienniczość w strukturę samej sieci neuronowej. Przykładem takim są omawiane dalej Konwolucyjne Sieci Neuronowe, których struktura jest skonstruowana w taki sposób, aby wyciągać pewne cechy przestrzenne z obrazów.

5.2 Bayesowskie Sieci Neuronowe (BNN)

5.3 Sieci konwolucyjne (CNN)

5.4 Sieci rekurencyjne (RNN)

5.5 Transformery

5.6 Normalizing flow

6 Uczenie ze wzmocnieniem i teoria optymalnego sterowania