

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

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

które spełnia: $P(\Omega) = 1$ oraz dla dowolnych parami rozłącznych zdarzeń $A_1, A_2, \dots \in \Sigma$ 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 $P(A \cup B) = P(A) + P(B) - P(A, B)$.

Prawdopodobieństwo warunkowe

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

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

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 $P(A | B) = P(A)$. Ponadto jeśli zdarzenia $A_1, A_2, \dots \in \Sigma$ są parami rozłączne i zachodzi $\bigcup_i A_i = \Omega$ to dla dowolnego zdarzenia $B \in \Sigma$ możemy zapisać

$$P(B) = \sum_i P(B | A_i)P(A_i).$$

Z definicji prawdopodobieństwa warunkowego trywialnie udowodnić twierdzenie Bayesa

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

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

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

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

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

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

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

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

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

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 $\mathbb{E}[f(\mathbf{x})] = F(t_1, \dots, t_n)$.

Rozkłady brzegowe

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

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

Zmienne losowe niezależne

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

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

Rozkłady warunkowe

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

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

Transformacja zmiennych wielowymiarowych

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

$$p_X(x_1, \dots, x_n) dx_1 \dots dx_n = p_Y(y_1, \dots, y_n) dy_1 \dots dy_n,$$

skąd

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

Macierz kowariancji

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

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

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, \dots, X_n są niezależne i f jest funkcją identycznościową to $\mathbf{\Sigma}$ jest macierzą diagonalną.

Wielowymiarowy rozkład normalny

Jeśli zmienna wielowymiarowa $X = (X_1, \dots, X_n)$ 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

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

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.

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

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

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

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

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

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

Pomiędzy tak zdefiniowanymi rodzajami zbieżności zachodzą następujące implikacje:

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

Wnioskowanie statystyczne

Modelem statystycznym nazwiemy parę (χ, \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, \dots, 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. $T = T(X_1, \dots, X_n)$. Być może najważniejszym przykładem statystyki jest średnia oznaczana jako \bar{X}

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

Silne prawo wielkich liczb

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

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

Wówczas zachodzi silne prawo wielkich liczb

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

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*.

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 (\bar{X}_n) ciąg średnich częściowych ciągu (X_n) zachodzi

$$\sqrt{n} (\bar{X}_n - \boldsymbol{\mu}) \rightarrow Z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

Oznacza to, iż dla ciągu X_1, \dots, 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 $\mathcal{N}(\boldsymbol{\mu}, n^{-1/2}\boldsymbol{\Sigma})$.

Estymatory punktowe MLE i MAP

Rozważamy model statystyczny $\mathcal{P} = \{p(\mathbf{x} | \theta) | \theta \in \Theta\}$. Estymatorem parametru θ nazwiemy statystykę $\hat{\theta}(X_1, \dots, X_n)$ służącą do oszacowania wartości tego parametru. Wartość tej statystyki dla konkretnej realizacji prostej próby losowej $\theta(\mathbf{x}_1, \dots, \mathbf{x}_n)$ 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 $\mathcal{P} = \{p(\mathbf{x} | \theta) | \theta \in \Theta\}$ i realizacji prostej próby losowej (którą nazwiemy również danymi lub obserwacjami) $D = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ jako

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

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

$$p(D | \theta_{\text{MLE}}) = \max_{\theta \in \Theta} p(D | \theta).$$

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

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

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

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

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 $\theta_{\text{MAP}} \in \Theta$, że

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

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

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

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 $D = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, wówczas twierdzenie Bayesa możemy zapisać jako

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

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, 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}

$$p(\mathbf{x} \mid D) = \int_{\Theta} p(\mathbf{x} \mid \theta) p(\theta \mid D) d\theta .$$

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

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

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

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

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 $(\mathbf{y}_1, \dots, \mathbf{y}_m)$. 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

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

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

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

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

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

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

$$\begin{aligned}p(\mathbf{x} \mid D) &\cong \pi(\mathbf{x}) \prod_{i=1}^m p(\mathbf{y}_i \mid \mathbf{x}) \cong \\ \exp \left\{ -\frac{1}{2} \left[(\mathbf{x} - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\mathbf{x} - \boldsymbol{\mu}_0) + \sum_{i=1}^m (\mathbf{y}_i - \mathbf{A}\mathbf{x} - \mathbf{b})^\top \boldsymbol{\Sigma}_y^{-1} (\mathbf{y}_i - \mathbf{A}\mathbf{x} - \mathbf{b}) \right] \right\}.\end{aligned}$$

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 $D = (y_1, \dots, y_m)$ dla zadanych (lub dokładnie znanych) przez nas $(\mathbf{x}_1, \dots, \mathbf{x}_m)$. Wiarygodność ma zatem postać

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

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

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

Człon stały możemy oczywiście pominąć i zapisać

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

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}_{MLE} musimy rozwiązać równanie

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

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 $p(\mathbf{w} | D)$. Jeśli jako prior przyjmujemy rozkład normalny z pewnymi parametrami $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$ to zauważmy, iż otrzymujemy instancję liniowego modelu Gaussowskiego

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

skąd rozkład a posteriori jest rozkładem normalnym

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

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

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

Nietrudno zauważyć, iż będzie to rozkład normalny o parametrach

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

oraz

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

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

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

Regularyzacja

Regularyzacją nazywamy proces polegający na wprowadzeniu ad hoc do zagadnienia optymalizacji dodatkowych członów tak, aby rozwiązanie było „reguluarne” (prostsze, nieosobliwe, jednoznaczne ...). W przypadku funkcji kosztu ℓ najczęściej dodajemy człon penalizujący rozwiązanie o dużej normie estymowanego parametru postaci

$$\gamma \|\boldsymbol{\theta}\|$$

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 przyjmujemy rozkład normalny

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

to rozkład a posteriori jest rozkładem normalnym o parametrach

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

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.

Procesy Gaussowskie

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

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

Pokazaliśmy również, iż macierz ta jest nieujemnie określona. Dodatkowo pokazaliśmy, 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

$$\mathbb{E}[(\mathbf{Lz})(\mathbf{Lz})^\top] = \mathbb{E}[\mathbf{Lzz}^\top \mathbf{L}^\top] = \mathbf{L}\mathbb{E}[\mathbf{zz}^\top]\mathbf{L}^\top = \mathbf{L}\mathbf{1}\mathbf{L}^\top = \mathbf{K}.$$

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 $\forall m \in \mathbb{N} : \forall X = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathbb{R}^n$ 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 $X = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathbb{R}^n$ i $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_s\} \subset \mathbb{R}^n$ 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 $X = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathbb{R}^n$ 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 proces 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 predykcji. Załóżmy, iż w zadanych (lub dokładnie znanych) przez nas punktach $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ 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 $Y = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_s\}$. 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

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

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

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

Wieloklasowa regresja logistyczna

Założmy, iż modelujemy obserwacje postaci (t, \mathbf{x}) , gdzie $t \in \{\tau_1, \tau_2, \dots, \tau_s\}$ to etykieta określająca przynależność do jednej z s klas, a $\mathbf{x} \in \mathbb{R}^n$ jest znanym (lub zadany) 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

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

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

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

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

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

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

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

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

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

Jako prior dla parametrów \mathbf{w}_j przyjmiemy rozkład normalny z pewnym hiperparametrem γ

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

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 zregulowanej funkcji kosztu

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

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

Metoda spadku wzdłuż gradientu

Algorytm spadku wzdłuż gradientu (z ang. *gradient descent*) dla funkcji $f(\mathbf{x}_1, \dots, \mathbf{x}_m)$ ma postać

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

2. Powtarzaj

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

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

Zakładając $\epsilon_1 = \dots = \epsilon_m = \epsilon$ 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

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

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

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

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

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

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 $\epsilon \leftarrow m^{-1}\epsilon$. Wówczas minimalizujemy tak naprawdę średni koszt ℓ^*/m .

Wnioskowanie metodami Monte Carlo