1 Wnioskowanie Bayesowskie

Typowym sposobem połączenia wyników matematycznej teorii prawdopodobieństwa z eksperymentami jest interpretacja częstościowa, która stwierdza iż prawdopodobieństwo zdarzenia A to częstość realizacji A w dużej liczbie prób. Dla wielu zdarzeń pojęcie ***dużej liczby prób*** nie jest jednak dobrze zdefiniowane. Wówczas interpretacja Bayesowska prawdopodobieństwa jako miary niepewności co do rezultatu określonego zdarzenia jest bardziej naturalna. Wnioskowanie Bayesowskie opiera się na twierdzeniu Bayesa

$$\boxed{\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}}\overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{\int p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}_{\text{evidence}}}}$$

oraz dwóch regułach znanych z rachunku prawdopodobieństwa: ***sum rule*** i ***product rule***.

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y})\,\mathrm{d}\boldsymbol{y}\ , \quad p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x} \mid \boldsymbol{y})p(\boldsymbol{y})\,.$$

Typowe zastosowanie wnioskowania Bayesowskie przebiega następująco: mamy pewien zbiór obserwacji $\mathcal{X}$, o których zakładamy, iż pochodzą z pewnego rozkładu prawdopodobieństwa z parametrami $\boldsymbol{\theta}$, tj. znamy wiarygodność $p(\mathcal{X} \mid \boldsymbol{\theta})$. Zakładamy dodatkowo pewien prior nad parametrami $p(\boldsymbol{\theta})$. Następnie korzystając z twierdzenia Bayesa wyznaczamy rozkład a posteriori nad parametrami modelu

$$p(\boldsymbol{\theta} \mid \mathcal{X}) = \frac{p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{X} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}}.$$

Rozkład a posteriori podsumowuje całą naszą wiedzę o estymowanym parametrze (z perspektywy wnioskowania Bayesowskiego). Korzystając z tego rozkładu możemy:

- Wyznaczyć estymatę punktową (tzw. estymatę ***maximum a posteriori*** (MAP)) parametru $\boldsymbol{\theta}$ jako
$$\boxed{\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathcal{X})} \quad .$$
- Wyznaczyć ***przedział wiarygodności*** (ang. ***credible interval***, nie mylić z przedziałem ufności) parametru $\boldsymbol{\theta}$ jako
$$C_\alpha(\boldsymbol{\theta}; \mathcal{X}) = (\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}})\,,$$
gdzie
$$\int_{\underline{\boldsymbol{\theta}}}^{\overline{\boldsymbol{\theta}}} p(\boldsymbol{\theta} \mid \mathcal{X})\,\mathrm{d}\boldsymbol{\theta} = 1 - \alpha\,.$$
- Skonstruować ***rozkład predyktywny*** (ang. ***posteriori predictive distribution***) dla nowych obserwacji
$$\boxed{p(\boldsymbol{x} \mid \mathcal{X}) = \int p(\boldsymbol{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathcal{X})\,\mathrm{d}\boldsymbol{\theta}} \quad .$$
- Możemy wyznaczyć wartość parametru $\boldsymbol{\theta}$, która minimalizuje oczekiwaną stratę dla pewnej funkcji kosztu
$$\begin{aligned}\boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}'} \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathcal{X})}\left[L(\boldsymbol{\theta}', \boldsymbol{\theta})\right] \\ &= \arg\min_{\boldsymbol{\theta}'} \int L(\boldsymbol{\theta}, \boldsymbol{\theta}')p(\boldsymbol{\theta} \mid \mathcal{X})\,\mathrm{d}\boldsymbol{\theta}\ .\end{aligned}$$

2 Rozkład normalny

Wielowymiarowy rozkład normalny o średniej $\boldsymbol{\mu} \in \mathbb{R}^n$ i symetrycznej, nieujemnie określonej macierzy kowariancji $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ to rozkład o gęstości danej przez

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right].$$

Macierz $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$ nazywamy macierzą precyzji. Będziemy często korzystać z następującej własności rozkładów normalnych: jeśli zmienne $\boldsymbol{x} \in \mathbb{R}^n$ i $\boldsymbol{y} \in \mathbb{R}^m$ mają łącznie wielowymiarowy rozkład normalny

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

to rozkłady brzegowe są rozkładami normalnymi

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

oraz rozkłady warunkowe są rozkładami normalnymi

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

gdzie

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

3 Liniowe modele Gaussowskie

Jednym z najprostszych modeli, do którego można zastosować wnioskowanie Bayesowskie jest tzw. liniowy model Gaussowski. Zakładamy, iż mamy obserwację $\boldsymbol{y}$, która pochodzi z rozkładu normalnego

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

gdzie $\boldsymbol{x}$ jest pewną zmienną ukrytą (ang. ***hidden variable***), natomiast $\boldsymbol{A}$, $\boldsymbol{b}$ i $\boldsymbol{\Sigma}_y$ są znane dokładnie. Dodatkowo przyjmijmy ***prior sprzężony do wiarygodności*** dla parametru $\boldsymbol{x}$ będący rozkładem normalnym o znanych parametrach $\boldsymbol{\mu}_x$, $\boldsymbol{\Sigma}_x$

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

Wówczas rozkład a posteriori również będzie rozkładem normalnym o parametrach

$$\begin{aligned}\boldsymbol{\mu}_{x|y} &= \boldsymbol{\Sigma}_{x|y}\left(\boldsymbol{A}^\intercal\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y}-\boldsymbol{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x\right)\\ \boldsymbol{\Sigma}_{x|y} &= \left(\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^\intercal\boldsymbol{\Sigma}_y^{-1}\boldsymbol{A}\right)^{-1}\end{aligned}\,.$$

4 Regresja liniowa

Załóżmy, że modelujemy obserwacje postaci $(y, \boldsymbol{x})$, gdzie y to skalar zwany ***zmienną objaśnianą***, natomiast $\boldsymbol{x}$ to wektor cech lub inaczej ***zmiennych objaśniających***. Zakładamy, że zmienna objaśniana zależy liniowo od zmiennych objaśniających oraz, iż zależność ta jest obarczona pewnym błędem losowym. Model ten ma więc postać

$$\boxed{y(\boldsymbol{x}) = \phi(\boldsymbol{x}; \boldsymbol{w}, b) + \epsilon = \boldsymbol{w}^\intercal\boldsymbol{x} + b + \epsilon\,,}$$

gdzie $y(\boldsymbol{x})$ jest wartością zmiennej losowej obserwowaną dla znanego dokładnie wektora cech $\boldsymbol{x}$ (wektor cech ***nie*** jest zmienną losową), natomiast $\boldsymbol{w}$ i b to parametry modelu ϕ, o których chcemy wnioskować. Model ten jest typowym przykładem ***regresji liniowej***. Parametryzację regresji liniowej można uprościć przyjmując, że do wektora cech dokładamy stałą wartość 1, a parametr b włączamy do $\boldsymbol{w}$

$$\boldsymbol{x}^\intercal \leftarrow \begin{bmatrix}\boldsymbol{x}^\intercal & 1\end{bmatrix}, \quad \boldsymbol{w}^\intercal \leftarrow \begin{bmatrix}\boldsymbol{w}^\intercal & b\end{bmatrix}.$$

W zależności od przyjętego rozkładu zmiennej opisującej błąd losowy ϵ otrzymujemy różne modele, np.:

- regresja liniowa (ang. ***linear regression***), $\epsilon \sim \mathcal{N}(0, \sigma^2)$ dla znanego σ^2,
- regresja odporna (ang. ***robust regression***), $\epsilon \sim \text{Laplace}(0, \lambda)$ dla znanego λ

$$\text{Laplace}(x; \mu, \lambda) = \frac{1}{2b}\exp\left[-\frac{|x-\mu|}{b}\right] \quad ,$$

- regresja kwantylowa (ang. ***quantile regression***), $\epsilon \sim \text{ALD}(0, \lambda, p)$ dla znanego λ i percentyla p

$$\text{ALD}(x; \mu, \lambda, p) = \frac{p(1-p)}{\lambda}\cdot\begin{cases}\mathrm{e}^{-\frac{(p-1)(x-\mu)}{\lambda}} & , x \le \mu\\ \mathrm{e}^{-\frac{p(x-\mu)}{\lambda}} & , x > \mu\end{cases}.$$

Zauważmy tutaj, że założenie iż znamy skalę (tj. parametry σ, λ) błędu jest dosyć mocne – z reguły nie wiemy jak dokładnie możemy przybliżyć zmienną objaśnianą. Dokładne wnioskowanie Bayesowskie przeprowadzimy tylko dla pierwszego z powyższych modeli, gdzie błędy mają rozkład normalny. To założenie, choć sensowne, ma również swoje konsekwencja – nasz model będzie dosyć wrażliwy na obserwacje odstające (ang. ***outliers***). Zamiast rozkładu normalnego można przyjąć rozkład, w którym gęstość prawdopodobieństwa ma tzw. ciężkie ogony (ang. ***heavy tails***) – np. rozkład Laplace'a. Otrzymamy wówczas model zwany odporną regresją liniową (ang. ***robust linear regression***). Oba powyższe modele równo rozkładają masę prawdopodobieństwa błędów. Jeśli interesuje nas natomiast estymacja kwantyli warunkowych (tj. prostej która najlepiej dzieli obserwacje tak aby odpowiedni ułamek z nich znalazł się pod nią), to możemy wykorzystać asymetryczny rozkład Laplace'a. Otrzymany model nazywamy wówczas kwantylową regresją liniową (ang. ***quantile linear regression***).

Przyjmijmy obecnie, że dysponujemy zbiorem obserwacji $\mathcal{X} = \{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. Zakładamy, że obserwacje te są i.i.d. – niezależne i pochodzące z tego samego rozkładu $y_i \sim \mathcal{N}(\phi(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2)$. Jawne wnioskowanie Bayesowskie możliwe jest również dla regresji liniowej. Bardzo często jednak stosuje się prostsze podejście adekwatne również do pozostałych dwóch modeli. Otóż poszukujemy jedynie estymaty punktowe parametrów $\boldsymbol{w}$, ignorując niepewność ich oszacowania. Taką estymację punktową można wyznaczyć ***metodą największej wiarygodności*** (ang. ***maximum likelihood estimation*** (MLE))

$$\boxed{\boldsymbol{w}_{\text{MLE}} = \arg\max_{\boldsymbol{w}} p(\mathcal{X} \mid \boldsymbol{w})}\,.$$

Korzystamy wówczas jedynie z wiarygodności $p(\mathcal{X} \mid \boldsymbol{w})$. Z założenia, iż obserwacje są i.i.d. możemy łatwo wyznaczyć wiarygodność

$$\begin{aligned}p(\mathcal{X} \mid \boldsymbol{w}) &= \prod_{i=1}^n \mathcal{N}(y_i; \phi(\boldsymbol{x}_i; \boldsymbol{w}), \sigma^2)\\ &\propto \prod_{i=1}^n \exp\left(-\frac{(y_i - \phi(\boldsymbol{x}_i; \boldsymbol{w}))^2}{2\sigma^2}\right),\end{aligned}$$

gdzie pominęliśmy współczynnik, gdyż nie ma on wpływu na zagadnienie optymalizacji. W praktyce często łatwiej jest minimalizować sumę pewnych wyrażeń, niż ich iloczyn, dlatego wprowadzamy ***zanegowaną logarytmiczną funkcję wiarygodności*** (ang. ***negated log-likelihood function*** (NLL))

$$\boxed{L(\mathcal{X}, \boldsymbol{w}) = -\log p(\mathcal{X} \mid \boldsymbol{w})},$$

gdzie korzystamy oczywiście z faktu, iż logarytm (naturalny) jest funkcją ściśle rosnącą. Estymatę MLE możemy wówczas znaleźć jako

$$\boxed{\boldsymbol{w}_{\text{MLE}} = \arg\min_{\boldsymbol{w}} L(\mathcal{X}, \boldsymbol{w})}.$$

W przypadku regresji liniowej otrzymujemy

$$\boldsymbol{w}_{\text{MLE}} = \arg\min_{\boldsymbol{w}} \frac{1}{2} \sum_{i=1}^{n} (y_i - \phi(\boldsymbol{x}_i; \boldsymbol{w}))^2 .$$

Z powyższego problemu optymalizacyjnego wynika popularna nazwa tej formy regresji liniowej – ***metoda najmniejszych kwadratów*** (ang. ***Ordinary Least Squares*** (OLS)), ponieważ minimalizujemy sumę kwadratów rezyduów. W tym szczególnym przypadku postać $\boldsymbol{w}_{\text{MLE}}$ można znaleźć analitycznie. Istotnie wprowadźmy

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^\intercal \\ \vdots \\ \boldsymbol{x}_n^\intercal \end{bmatrix}, \quad \boldsymbol{\phi} = \boldsymbol{X}\boldsymbol{w} .$$

Wówczas mamy

$$\frac{\partial L}{\partial \phi_\beta} = (\phi_\beta - y_\beta)$$

oraz dalej

$$\frac{\partial L}{\partial w_\alpha} = \sum_{\beta=1}^{n} \frac{\partial L}{\partial \phi_\beta} \frac{\partial \phi_\beta}{\partial w_\alpha} = \sum_{\beta=1}^{n} (\phi_\beta - y_\beta) X_{\beta\alpha} .$$

Z warunków koniecznych dla minimum lokalnego mamy $\forall \alpha : \frac{\partial L}{\partial w_\alpha} = 0$, co możemy zapisać jako

$$\boldsymbol{X}^\intercal(\boldsymbol{\phi} - \boldsymbol{y}) = \boldsymbol{X}^\intercal(\boldsymbol{X}\boldsymbol{w}_{\text{MLE}} - \boldsymbol{y}) = \boldsymbol{0} ,$$

skąd

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

Wielkość $(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal$ nazywa się pseudo odwrotnością Moore'a–Penrose macierzy $\boldsymbol{X}$. Warto tutaj zaznaczyć, iż w praktyce rzadko wyznaczamy $\boldsymbol{w}_{\text{MLE}}$ korzystając wprost z powyższego wzoru, ze względu na dość słabe własności numeryczne. Typowo jeśli macierz $\boldsymbol{X}$ jest niewielka, to pseudo odwrotność obliczamy korzystając z rozkładu SVD macierzy $\boldsymbol{X}$. Natomiast dla dużych macierzy nie korzystamy ze wzoru analitycznego tylko z metody iteracyjnej stochastycznego spadku wzdłuż gradientu

$$w_\alpha^{(t+1)} = w_\alpha^{(t)} - \eta \frac{\partial L}{\partial w_\alpha} ,$$

która w tym przypadku zbiega do minimum globalnego $\boldsymbol{w}_{\text{MLE}}$.

4.1 Regresja Bayesowska

Pokażemy teraz, jak przeprowadzić pełne wnioskowanie Bayesowskie dla modelu regresji liniowej. Założymy rozkład a priori nad parametrami $\boldsymbol{w}$ w postaci rozkładu normalnego

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) .$$

Wiarygodność, jak pokazaliśmy powyżej ma natomiast postać

$$\mathcal{X} \mid \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{w}, \sigma^2 \boldsymbol{I}) .$$

Zauważmy, że jest to instancja modelu liniowego, więc korzystając z wyprowadzonych wzorów wiemy, iż

$$\boldsymbol{w} \mid \mathcal{X} \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) ,$$

gdzie

$$\begin{aligned} \boldsymbol{\mu}_n &= \boldsymbol{\Sigma}_n \left(\frac{1}{\sigma^2} \boldsymbol{X}^\intercal \boldsymbol{y} + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right) \\ \boldsymbol{\Sigma}_n &= \left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{X}^\intercal \boldsymbol{X} \right)^{-1} \end{aligned} .$$

W powyższych wzorach nazwy parametrów nie są przypadkowe. Przed zaobserwowaniem danych mamy rozkład a priori $\boldsymbol{w} \mid \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Po zaobserwowaniu n przykładów aktualizujemy nasze przekonania w postaci rozkładu a posteriori $\boldsymbol{w} \mid \mathcal{X} \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$.

Znając rozkład a posteriori nad $\boldsymbol{w}$ możemy również wyznaczyć rozkład predykcyjny nad zmienną objaśnianą y dla ustalonego wektora cech $\boldsymbol{x}$. Formalnie interesuje nas rozkład $p(y \mid \mathcal{X})$. Rozkład ten możemy opisać w następujący sposób

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

gdzie interesuje nas rozkład

$$p(y \mid \mathcal{X}) = \int p(y \mid \boldsymbol{w}) p(\boldsymbol{w} \mid \mathcal{X}) \, \mathrm{d}\boldsymbol{w} .$$

Można pokazać, iż rozkład ten jest rozkładem normalnym

$$y \mid \mathcal{X} \sim \mathcal{N}(\mu_y, \sigma_y^2)$$

o parametrach

$$\mu_y = \boldsymbol{\mu}_n^\intercal \boldsymbol{x} , \quad \sigma_y^2 = \sigma^2 + \boldsymbol{x}^\intercal \boldsymbol{\Sigma}_n \boldsymbol{x} .$$

4.2 Regresja grzbietowa

Rozważmy Bayesowską regresję liniową, w której prior ma postać $\boldsymbol{w} \sim \mathcal{N}(0, \tau^2 \boldsymbol{I})$, czyli elementy wektora parametrów są a priori wzajemnie niezależne a ich amplitudy są rzędu τ. Znajdźmy estymatę MAP parametrów $\boldsymbol{w}$. Zauważmy, że

$$-\log p(\boldsymbol{w} \mid \mathcal{X}) \propto \left[\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \phi(\boldsymbol{x}_i; \boldsymbol{w}))^2 + \frac{1}{2\tau^2} \|\boldsymbol{w}\|_2^2 \right].$$

Wyrażenie w nawiasach możemy przemnożyć przez σ^2 – nie wpłynie to na położenie minimum, bo σ^2 to dodatnia stała. Estymata MAP w tym modelu to zatem

$$\boldsymbol{w}_{\mathrm{MAP}} = \arg\min_{\boldsymbol{w}} \left[\frac{1}{2} \sum_{i=1}^{n} (y_i - \phi(\boldsymbol{x}_i; \boldsymbol{w}))^2 + \frac{\gamma}{2} \|\boldsymbol{w}\|_2^2 \right]$$

gdzie $\gamma = \sigma^2/\tau^2$ to ***stała regularyzująca***. Pierwszy z członów to znana z OLS suma kwadratów rezyduów, natomiast drugi człon to ***człon regularyzujący*** rozwiązanie, który ogranicza amplitudy parametrów (wag). Jest to tzw. model ***regresji grzbietowej*** (ang. ***ridge regression***). Jest to jednocześnie oczywiście moda rozkładu a posteriori, czyli w przypadku rozkładu normalnego – wyznaczony wcześniej parametr $\boldsymbol{\mu_n}$

$$\boldsymbol{w}_{\mathrm{MAP}} = \left(\frac{\sigma^2}{\tau^2} \boldsymbol{I} + \boldsymbol{X}^\intercal \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}\,.$$

5 Wprowadzenie do procesów Gaussowskich

Macierz kowariancji n–wymiarowej zmiennej losowej $\boldsymbol{x}$ o wartości oczekiwanej $\boldsymbol{\mu}$ jest zdefiniowana jako

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

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

$$\boldsymbol{K} = \boldsymbol{L}\boldsymbol{L}^\intercal,$$

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

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

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

Funkcję $k : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ taką, że $\forall m \in \mathbb{N} : \forall X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\} \subset \mathbb{R}^n$ macierz

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

jest dodatnio określoną macierzą symetryczną nazywamy funkcją kowariancji, jądrem dodatnio określonym (ang. ***positive definite kernel***) lub ***jądrem Mercera***.

Dla dwóch zbiorów punktów $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\} \subset \mathbb{R}^n$ i $Y = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_s\} \subset \mathbb{R}^n$ i funkcji kowariancji k wprowadzimy oznaczenie

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

Poniżej podajemy kilka przykładów funkcji kowariancji

- ***Gaussian kernel*** dla normy $\|\cdot\|$ i hiperparametrów a, l (amplituda i skala długości)

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

- ***Periodic kernel*** dla normy $\|\cdot\|$ i hiperparametrów a, l, p (amplituda, skala długości, okres zmienności)

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

- ***White noise kernel*** dla hiper-parametru σ

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

- ***Matérn kernel*** dla normy $\|\cdot\|$ i hiperparametrów a, l, ν (amplituda, skala długości, regularność)

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

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

Suma lub iloczyn dwóch funkcji kowariancji oraz złożenie funkcji kowariancji z wielomianem o nieujemnych współczynnikach jest również funkcją kowariancji

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

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

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

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

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 $\boldsymbol{\mu}_X = \boldsymbol{0}$, natomiast przepisem na macierz kowariancji może być zdefiniowana wyżej funkcja kowariancji $k(X, X)$ tj.

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

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 dokładnie znanych przez nas punktach $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$ 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

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

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

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

Zauważmy, iż z twierdzenia o własnościach niezdegenerowanego rozkładu normalnego wnioskujemy, iż warunkowy $\boldsymbol{f}_Y \mid \boldsymbol{f}_X$ jest również rozkładem normalnym o parametrach

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

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

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

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

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

6 Regresja logistyczna

Rozpatrzymy teraz problem klasyfikacji binarnej z perspektywy estymacji MLE. Niech $\mathcal{X} = \{t_i(\boldsymbol{x}_i)\}_{i=1}^n$ będzie zbiorem obserwacji i.i.d., gdzie zakładamy, iż $t \in \{0, 1\}$. Jako model statystyczny przyjmiemy, iż klasa $t(\boldsymbol{x})$ pochodzi z rozkładu Bernoulliego z parametrem $\pi(\boldsymbol{x}; \boldsymbol{w})$ (prawdopodobieństwem klasy pozytywnej, gdzie π jest tzw. funkcją wiążącą, ang. ***link function***) zależnym od estymowanych parametrów $\boldsymbol{w}$

$$t_i \mid \boldsymbol{w} \sim \mathrm{Ber}(\pi(\boldsymbol{x}_i; \boldsymbol{w}))\,.$$

W przypadku regresji logistycznej jako funkcję $\pi(\boldsymbol{x}; \boldsymbol{w})$ przyjmiemy $\sigma(\phi(\boldsymbol{x}; \boldsymbol{w}))$, gdzie

$$\sigma(z) = \frac{1}{1 + \mathrm{e}^{-z}}$$

to tzw. ***funkcja logistyczna*** (będąca szczególnym przypadkiem funkcji sigmoidalnej), natomiast $\phi(\boldsymbol{x}; \boldsymbol{w})$ może być dowolną funkcją wykorzystywaną do zagadnienia regresji. W przypadku regresji logistycznej przyjmujemy prosty model liniowy

$$\phi(\boldsymbol{x}; \boldsymbol{w}) = \boldsymbol{w}^\intercal\boldsymbol{x}\,.$$

Zauważmy, iż taka postać uwzględnia człon stały poprzez podstawienie $\boldsymbol{x}^\intercal \leftarrow \begin{bmatrix}\boldsymbol{x}^\intercal & 1\end{bmatrix}$. Wiarygodność dla powyższego modelu statystycznego ma postać

$$p(\mathcal{X} \mid \boldsymbol{w}) = \prod_{i=1}^n \pi(\boldsymbol{x}_i; \boldsymbol{w})^{t_i}(1 - \pi(\boldsymbol{x}_i; \boldsymbol{w}))^{1-t_i}\,,$$

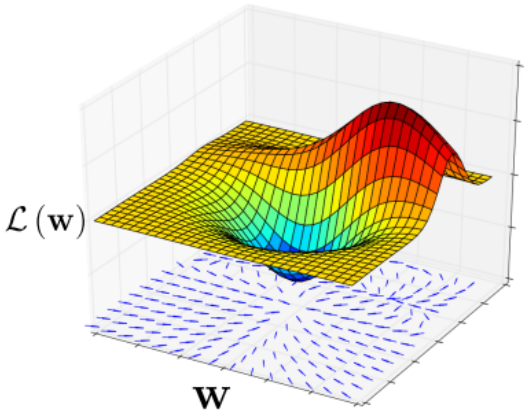
skąd funkcja NLL

$$\boxed{L(\mathcal{X};\boldsymbol{w}) = -\sum_{i=1}^{n} \left[t_i \log \pi_i + (1-t_i)\log(1-\pi_i)\right]\,,}$$

gdzie $\pi_i = \pi(\boldsymbol{x}_i;\boldsymbol{w})$. Taką funkcję kosztu nazywamy ***binarną entropią krzyżową*** (ang. ***binary cross-entropy function***). Niestety dla takiej postaci funkcji kosztu nie można znaleźć minimum w postaci analitycznej (jak zrobiliśmy w przypadku regresji liniowej), dlatego musimy wykorzystywać algorytmy optymalizacji numerycznej, które najczęściej wykorzystują pierwsze pochodne funkcji kosztu po parametrach. Typowo wykorzystywanym do tego celu algorytmem jest ***spadek wzdłuż gradientu*** (ang. ***gradient descent***), w którym iteracyjnie wykonujemy

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \frac{\partial L}{\partial \boldsymbol{w}}\bigg|_{\boldsymbol{w}_t}\,.$$

Wyznaczymy więc pochodne funkcji L po parame-



trach $\boldsymbol{w}$. Wprowadźmy macierz $X_{\alpha\beta} = x_\beta^{(\alpha)}$ oraz wektory $\phi_\alpha = \phi(\boldsymbol{x}_\alpha;\boldsymbol{w})$, $\sigma_\alpha = \sigma(\phi_\alpha)$. Wówczas

$$L(\mathcal{X};\boldsymbol{w}) = -\sum_{\alpha=1}^{n} \left[t_\alpha \log \sigma_\alpha + (1-t_\alpha)\log(1-\sigma_\alpha)\right]\,.$$

W takim razie mamy

$$\frac{\partial L}{\partial \phi_\alpha} = \sum_{\beta=1}^{n} \frac{\partial L}{\partial \sigma_\beta}\frac{\partial \sigma_\beta}{\partial \phi_\alpha}\,,$$

gdzie

$$\frac{\partial L}{\partial \sigma_\beta} = \frac{1-t_\beta}{1-\sigma_\beta} - \frac{t_\beta}{\sigma_\beta}$$

oraz

$$\frac{\partial \sigma_\beta}{\partial \phi_\alpha} = \sigma'(\phi_\beta)\delta_{\alpha\beta}\,,$$

gdzie

$$\sigma'(z) = \frac{\mathrm{e}^{-z}}{(1+\mathrm{e}^{-z})^2} = \sigma(z)(1-\sigma(z))\,,$$

zatem

$$\frac{\partial L}{\partial \phi_\alpha} = \left(\frac{1-t_\alpha}{1-\sigma_\alpha} - \frac{t_\alpha}{\sigma_\alpha}\right)\sigma_\alpha(1-\sigma_\alpha) = \sigma_\alpha - t_\alpha\,.$$

W takim razie, z powyższego mamy

$$\frac{\partial L}{\partial w_\alpha} = \sum_{\beta=1}^{n} \frac{\partial L}{\partial \phi_\beta}\frac{\partial \phi_\beta}{\partial w_\alpha} = \sum_{\beta=1}^{n} (\sigma_\beta - t_\beta)X_{\beta\alpha}\,,$$

co możemy zapisać w zwartej postaci macierzowej

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{X}^\intercal(\boldsymbol{\sigma} - \boldsymbol{t})\,.$$

7 Wieloklasowa regresja logistyczna

Regresję logistyczną możemy w prosty sposób uogólnić na przypadek klasyfikacji jednej z k klas. Zakładamy teraz, iż zbiór obserwacji i.i.d. ma postać $\mathcal{X} = \{t_i(\boldsymbol{x}_i)\}_{i=1}^n$ dla $t \in \{1,\ldots,k\}$. Nasz model statystyczny uogólnimy do postaci rozkładu kategorialnego (ang. ***categorical distribution*** lub ***multinomial distribution***) z parametrem $\boldsymbol{\pi}(\boldsymbol{x};\boldsymbol{W})$ (wektorem prawdopodobieństw każdej klasy) zależnym od estymowanych parametrów $\boldsymbol{W}$

$$t_i \mid \boldsymbol{W} \sim \mathrm{Cat}(\boldsymbol{\pi}(\boldsymbol{x}_i;\boldsymbol{W}))\,.$$

W przypadku regresji softmax jako funkcję $\boldsymbol{\pi} : \mathbb{R}^m \mapsto [0;1]^k$ przyjmiemy funkcję $\boldsymbol{\sigma}(\boldsymbol{\phi}(\boldsymbol{x};\boldsymbol{W}))$, gdzie $\boldsymbol{\sigma} : \mathbb{R}^k \mapsto [0;1]^k$

$$\sigma_\alpha(\boldsymbol{z}) = \frac{\exp(z_\alpha)}{\sum_{\beta=1}^k \exp(z_\beta)}$$

to tzw. ***funkcja softmax***, natomiast $\boldsymbol{\phi} : \mathbb{R}^m \mapsto \mathbb{R}^k$, to dowolna funkcja. W przypadku regresji softmax przyjmiemy prosty model liniowy

$$\boldsymbol{\phi}(\boldsymbol{x};\boldsymbol{W}) = \boldsymbol{W}\boldsymbol{x}\,.$$

Wiarygodność dla powyższego modelu ma postać

$$\prod_{\alpha=1}^{n}\prod_{\beta=1}^{k} \pi_{\alpha\beta}^{t_{\alpha\beta}}\,,$$

skąd funkcja NLL ma postać

$$\boxed{L(\mathcal{X};\boldsymbol{W}) = -\sum_{\alpha=1}^{n}\sum_{\beta=1}^{k} t_{\alpha\beta}\log \pi_{\alpha\beta}\,,}$$

gdzie

$$\pi_{\alpha\beta} = \sigma_\beta(\boldsymbol{\phi}(\boldsymbol{x}_\alpha; \boldsymbol{W})) = \frac{\exp \phi_{\alpha\beta}}{\sum_{\gamma=1}^{k} \exp \phi_{\alpha\gamma}} .$$

$$\phi_{\alpha\beta} = \phi_\beta(\boldsymbol{x}_\alpha; \boldsymbol{W})$$

$$t_{\alpha\beta} = [t_\alpha = \beta]$$

Wiersze macierzy $\boldsymbol{t}$ są binarnymi wektorami kodującymi w sposób one-hot klasę danego przykładu. Niestety dla takiej postaci funkcji kosztu nie można znaleźć minimum w postaci analitycznej, dlatego musimy wykorzystywać algorytmy optymalizacji numerycznej. Wyznaczymy więc jeszcze pierwsze pochodne funkcji L po parametrach $\boldsymbol{W}$

$$\frac{\partial L}{\partial \pi_{\beta\beta'}} = -\frac{t_{\beta\beta'}}{\pi_{\beta\beta'}} .$$

Jednocześnie mamy

$$\frac{\partial L}{\partial \phi_{\alpha\alpha'}} = \sum_{\beta=1}^{n} \sum_{\beta'=1}^{k} \frac{\partial L}{\partial \pi_{\beta\beta'}} \frac{\partial \pi_{\beta\beta'}}{\partial \phi_{\alpha\alpha'}} ,$$

gdzie

$$\frac{\partial \pi_{\beta\beta'}}{\partial \phi_{\alpha\alpha'}} = \frac{\exp \phi_{\beta\beta'} \delta_{\alpha\beta} \delta_{\alpha'\beta'}}{\sum_{\gamma=1}^{k} \exp \phi_{\beta\gamma}} - \frac{\exp \phi_{\beta\beta'} \exp \phi_{\beta\alpha'} \delta_{\alpha\beta}}{\left[\sum_{\gamma=1}^{k} \exp \phi_{\beta\gamma}\right]^2}$$

$$= \pi_{\beta\beta'} \delta_{\alpha\beta} \delta_{\alpha'\beta'} - \pi_{\beta\beta'} \pi_{\beta\alpha'} \delta_{\alpha\beta} .$$

Z powyższego mamy zatem

$$\frac{\partial L}{\partial \phi_{\alpha\alpha'}} = \pi_{\alpha\alpha'} \sum_{\beta'=1}^{k} t_{\alpha\beta'} - t_{\alpha\alpha'} = \pi_{\alpha\alpha'} - t_{\alpha\alpha'} .$$

Ostatecznie zatem

$$\frac{\partial L}{\partial W_{\beta\beta'}} = \sum_{\alpha=1}^{n} \sum_{\alpha'=1}^{k} \frac{\partial L}{\partial \phi_{\alpha\alpha'}} \frac{\partial \phi_{\alpha\alpha'}}{\partial W_{\beta\beta'}} ,$$

gdzie

$$\frac{\partial \phi_{\alpha\alpha'}}{\partial W_{\beta\beta'}} = X_{\alpha\beta'} \delta_{\alpha'\beta} ,$$

zatem

$$\frac{\partial L}{\partial W_{\beta\beta'}} = \sum_{\alpha=1}^{n} (\pi_{\alpha\beta} - t_{\alpha\beta}) X_{\alpha\beta'} ,$$

co możemy zapisać w zwartej postaci macierzowej

$$\frac{\partial L}{\partial \boldsymbol{W}} = (\boldsymbol{\pi} - \boldsymbol{t})^\intercal \boldsymbol{X} .$$

8 Liczby losowe w komputerze

Opiszemy teraz pokrótce metody generowania liczb pseudolosowych z dowolnych rozkładów prawdopodobieństwa w sposób algorytmiczny. Podstawowym narzędziem, którego będziemy potrzebować do generowania próbek z bardziej skomplikowanych rozkładów będzie prosty generator liczb z rozkładu jednostajnego $\mathcal{U}(0, 1)$. Moglibyśmy oczywiście wykorzystać jakieś fizyczne urządzenie lub proces, który generuje liczby prawdziwie losowe (np. detektor Geigera-Mullera, szum lamp elektronowych, ruletka), ale błędem byłaby rezygnacja z odtwarzalności. Poszukujemy zatem deterministycznej metody, która generuje sekwencje liczb, które są w przybliżeniu losowe. Podstawową metodą do algorytmicznego generowania liczb pseudolosowych jest tzw. ***liniowy generator kongruentny*** (ang. ***Linear Congruential Generator, LCG***), który jest opisany zależnością rekurencyjną

$$I_{j+1} = (aI_j + c) \mod m ,$$

gdzie a, c, m to pewne ustalone dodatnie liczby całkowite, a I_0 to tzw. ziarno (ang. ***seed***). LCG generuje liczby całkowite, więc w dużym uproszczeniu liczby zmiennoprzecinkowe z rozkładu $\mathcal{U}(0, 1)$ otrzymujemy jako I_j/m (trzeba tutaj jednak uwzględnić problemy wynikające z arytmetyki zmiennoprzecinkowej).

8.1 Metoda odwrotnej dystrybuanty

Mając już generator liczb z rozkładu jednostajnego $\mathcal{U}(0, 1)$ i znając jawny wzór na dystrybuantę $F(x)$ innego rozkładu jednowymiarowego możemy generować liczby z tego rozkładu korzystając z tzw. ***metody odwrotnej dystrybuanty***. Istotnie, jeśli $U \sim \mathcal{U}(0, 1)$, to $F^{-1}(U) \sim F$. Istotnie

$$\Pr\big(F^{-1}(U) \le x\big) = \Pr(U \le F(x)) = F(x) .$$

Metoda ta ma jedną zasadniczą wadę – musimy znać jawny wzór na dystrybuantę F. W przypadku np. tak ważnych rozkładów, jak rozkład normalny dystrybuanta nie jest funkcją elementarną i metoda ta nie jest najlepsza.

8.2 Metoda Boxa–Mullera

W przypadku rozkładu normalnego znacznie lepszą metodą jest tzw. ***metoda Boxa–Mullera***. Weźmy rozkład łączny dwóch niezależnych zmiennych losowych X, Y pochodzących ze standardowego

rozkładu normalnego

$$p(x,y) = \frac{1}{2\pi}\exp\left(\frac{1}{2}(x^2+y^2)\right).$$

Skorzystamy ze wzoru na transformację zmiennych losowych. Istotnie niech

$$X = \sqrt{Z}\cos\Phi\,, \quad Y = \sqrt{Z}\sin\Phi\,,$$

dla $0 < Z$ oraz $0 \le \Phi < 2\pi$, wówczas

$$q(z,\phi) = \frac{1}{4\pi}\exp\frac{z}{2}\,.$$

Zauważmy, że otrzymaliśmy sferycznie symetryczny rozkład wykładniczy. Możemy zatem wylosować kąt ϕ z rozkładu jednostajnego oraz wartość z z rozkładu wykładniczego korzystając z metody odwrotnej dystrybuanty. Wówczas wartości $x = \sqrt{z}\cos\phi$, $y = \sqrt{z}\sin\phi$ będą pochodzić ze standardowego rozkładu normalnego. Aby wygenerować próbki z ogólnego wielowymiarowego rozkładu normalnego, korzystamy z definicji, tj. najpierw generujemy n próbek ze standardowego rozkładu normalnego, a następnie korzystamy z przekształcenia afinicznego $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu}$, gdzie $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^\intercal$.

8.3 Monte Carlo

Umiemy już generować próbki z wielowymiarowego rozkładu normalnego. Chcemy teraz poznać metodę, która umożliwi generowanie próbek ze skomplikowanych, wielowymiarowych rozkładów prawdopodobieństwa, których gęstość znamy jedynie z dokładnością do stałej normalizującej, tj. znamy jedynie $\tilde{p}(\boldsymbol{x}) = Zp(\boldsymbol{x})$. Ograniczenie to wynika z chęci próbkowania z posteriora $p(\boldsymbol{x} \mid \boldsymbol{y})$ w sytuacji, gdy znamy jedynie rozkład łączny $p(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x})p_{\boldsymbol{X}}(\boldsymbol{x})$. Okazuje się, iż znajomość rozkładu jedynie z dokładnością do stałej normalizującej jest wystarczająca do generowania próbek z tego rozkładu. Generowanie próbek z kolei wystarcza natomiast, na mocy silnego prawa wielkich liczb, do szacowania wartości średnich dowolnych funkcji zmiennej $\boldsymbol{x}$. Przypomnijmy, iż na mocy silnego prawa wielkich liczb ciąg średnich częściowych $(\overline{\boldsymbol{X}}_n)$ ciągu zmiennych losowych $(\boldsymbol{X}_n)$ i.i.d. z rozkładu $\boldsymbol{X} \sim \mathcal{D}$ jest zbieżny z prawdopodobieństwem 1 do wartości oczekiwanej $\mathbb{E}[\boldsymbol{X}]$ tj.

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

Wartość oczekiwaną $\mathbb{E}[\boldsymbol{X}]$ możemy zatem przybliżyć średnią $\overline{\boldsymbol{X}}_n$ z dużej ilości próbek. Poniżej przedstawimy dwa algorytmy próbkowania: algorytm IS oraz Metropolisa–Hastingsa będący szczególną realizacją całej rodziny algorytmów próbkowania zwanych Markov Chain Monte Carlo (MCMC).

8.3.1 Importance sampling

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

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

tj. szukamy

$$\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[f(\boldsymbol{x})] = \int f(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\,.$$

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

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[f(\boldsymbol{x})] &= \int f(\boldsymbol{x})p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int f(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}\left[f(\boldsymbol{x})\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right] \\
&= \frac{Z_q}{Z_p}\mathbb{E}_q\left[f(\boldsymbol{x})\frac{\tilde{p}(\boldsymbol{x})}{\tilde{q}(\boldsymbol{x})}\right].
\end{aligned}$$

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

$$Z_p = \int \tilde{p}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = Z_q\int\frac{\tilde{p}(\boldsymbol{x})}{\tilde{q}(\boldsymbol{x})}q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = Z_q\mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}\left[\frac{\tilde{p}(\boldsymbol{x})}{\tilde{q}(\boldsymbol{x})}\right],$$

skąd ostatecznie

$$\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[f(\boldsymbol{x})] = \frac{\mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}\left[f(\boldsymbol{x})\frac{\tilde{p}(\boldsymbol{x})}{\tilde{q}(\boldsymbol{x})}\right]}{\mathbb{E}_{\boldsymbol{x}\sim q(\boldsymbol{x})}\left[\frac{\tilde{p}(\boldsymbol{x})}{\tilde{q}(\boldsymbol{x})}\right]}\,.$$

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

$$\boxed{\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}[f(\boldsymbol{x})] \approx \frac{\sum_{i=1}^m f(\boldsymbol{x}_i)\frac{\tilde{p}(\boldsymbol{x}_i)}{\tilde{q}(\boldsymbol{x}_i)}}{\sum_{j=1}^m \frac{\tilde{p}(\boldsymbol{x}_j)}{\tilde{q}(\boldsymbol{x}_j)}} = \sum_{i=1}^m \lambda_i f(\boldsymbol{x}_i)\,,}$$

gdzie

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

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

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

dla $\boldsymbol{x}_i \sim q$ mamy

$$\begin{aligned} \mathbb{V}[\overline{f}_m] &= \frac{1}{m} \mathbb{V}_q \left[f(\boldsymbol{x}) \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \right] \\ &= \frac{1}{m} \int \frac{(f(\boldsymbol{x}) p(\boldsymbol{x}) - \mu_f q(\boldsymbol{x}))^2}{q(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x} \, . \end{aligned}$$

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 $f(\boldsymbol{x})p(\boldsymbol{x})$, co może być trudne do praktycznego zrealizowania.

8.3.2 Algorytm Metropolisa–Hastingsa

Cała klasa algorytmów próbkowania MCMC opiera się na idei wyrażenia generowania próbek jako ewolucji pewnego łańcucha Markowa.

Łańcuchem Markowa nazwiemy ciąg zmiennych losowych $(\boldsymbol{X}_t)$ o wartościach w $\mathbb{R}^n$ taki, że spełnione jest ***kryterium Markowa***

$$\begin{aligned} \forall A \subset \mathbb{R}^n : \Pr(\boldsymbol{X}_t \in A \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1}, \ldots, \boldsymbol{X}_0 = \boldsymbol{x}_0) \\ = \Pr(\boldsymbol{X}_t \in A \mid \boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1}) \, . \end{aligned}$$

Elementy ciągu nazywamy stanami łańcucha.

Dany łańcuch jest zadany jednoznacznie przez podanie gęstości prawdopodobieństwa przejścia łańcucha ze stanu $\boldsymbol{x} \to \boldsymbol{y}$, którą będziemy oznaczać przez $\pi(\boldsymbol{y} \mid \boldsymbol{x})$ (zakładamy, iż prawdopodobieństwo przejścia jest niezależne od chwili t – łańcuch taki nazywamy jednorodnym). Funkcja π spełnia oczywiście warunek unormowania

$$\int \pi(\boldsymbol{y} \mid \boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} = 1 \, ,$$

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

$$p(\boldsymbol{y}) = \int \pi(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, .$$

Rozkład stacjonarny danego łańcucha oznaczymy przez $p^*(\boldsymbol{x})$. Zauważmy, iż jeśli stan początkowy łańcucha $\boldsymbol{X}_0$ pochodzi z rozkładu stacjonarnego p^* to każdy kolejny stan $\boldsymbol{X}_t$ 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

$$p_t(\boldsymbol{y}) = \int \pi(\boldsymbol{y} \mid \boldsymbol{x}) p_{t-1}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, , \quad \text{dla } t > 1.$$

Rozkładem granicznym łańcucha Markowa nazwiemy granicę w sensie zbieżności punktowej

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

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 $p(\boldsymbol{x})$ był rozkładem stacjonarnym łańcucha Markowa jest

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

co wynika z scałkowania powyższego równania. Kryterium to nazywamy ***kryterium lokalnego balansu*** (ang. ***detailed balance condition***).

Podstawowa idea wykorzystania łańcuchów Markowa do generowania próbek ze skomplikowanego rozkładu p jest więc następująca: tworzymy łańcuch Markowa, dla którego p jest rozkładem stacjonarnym, wówczas rozpoczynając w dowolnym dopuszczalnym stanie początkowym $\boldsymbol{X}_0$ po wykonaniu dużej liczby kroków (etap ten nazywamy okresem przejściowym ang. ***burn-in period***) stan $\boldsymbol{X}_t$ (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 $\pi(\boldsymbol{y} \mid \boldsymbol{x})$ dla danego rozkładu granicznego $p(\boldsymbol{x})$. Podstawową konstrukcję podaje algorytm Metropolisa–Hastingsa.

```
1  # Metropolis-Hastings
2  # --------------------
3  # Initialize state x using any admissible value.
4
5  for i = 1,...,max_iter:
6
7      # Step 1. Sample candidate y from the
8      # conditional proposal distribution q(y | x).
```

```
9
10      # Step 2. Accept new candidate with
11      # probability
12      r(y | x) = min{1, p(y)q(x|y) / p(x)q(y|x)}
13      # and change state to y, otherwise stay
14      # in state x.
```

Funkcja przejścia ma postać

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

gdzie

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

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

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

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

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

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***).

9 (Histogram) Gradient Boosted Trees

10 KDE

W poprzednim paragrafie opisaliśmy w jaki sposób mając (nieznormalizowaną) funkcję gęstości prawdopodobieństwa $\tilde{p}(\boldsymbol{x})$ generować algorytmicznie próbki z opisanego przez nią rozkładu. Teraz zajmiemy się problemem odwrotnym tj. mając realizację prostej próby losowej z pewnego rozkładu chcemy znaleźć funkcję $\hat{p}(\boldsymbol{x})$, która estymuje gęstość rozkładu prawdopodobieństwa, z którego pochodzą próbki. Opiszemy tutaj jedną z najprostszych metod zwaną estymatorem jądrowym (ang. ***Kernel Density Estimator, KDE***). Estymatorem jądrowym gęstości funkcji p nazywamy funkcję

$$\hat{p}(\boldsymbol{x}) := \frac{1}{n} \sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{x}_i) \,,$$

gdzie $\boldsymbol{H}$ jest symetryczną i dodatnio określoną macierzą zwaną *bandwidth matrix*, funkcja $K_{\boldsymbol{H}}$ ma postać

$$K_{\boldsymbol{H}}(\boldsymbol{x}) = (\det \boldsymbol{H})^{-1/2} K(\boldsymbol{H}^{-1/2} \boldsymbol{x}) \,,$$

gdzie funkcja K zwana jądrem (z ang. *kernel*) jest gęstością prawdopodobieństwa pewnego sferycznie symetrycznego rozkładu wielowymiarowego. Wybór funkcji K nie jest kluczowy ze statystycznego punktu widzenia, więc możemy bez problemu założyć, iż jest to gęstość wielowymiarowego rozkładu normalnego, tj.

$$K_{\boldsymbol{H}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^m \det \boldsymbol{H}}} \exp\left(-\frac{1}{2} \boldsymbol{x}^T \boldsymbol{H}^{-1} \boldsymbol{x}\right) ,$$

gdzie zakładamy $\boldsymbol{x} \in \mathbb{R}^m$. Większym problemem w przypadku KDE jest wybór odpowiedniego parametru $\boldsymbol{H}$. Jednym z prostszych wyborów w przypadku estymatora jest tzw. ***reguła Silvermana***, która podaje następujący przepis na macierz $\boldsymbol{H}$

$$H_{\alpha\beta} = 0 \,, \quad \sqrt{H_{\alpha\alpha}} = \left(\frac{4}{4+m}\right)^{\frac{1}{m+4}} n^{\frac{-1}{m+4}} \sigma_\alpha \,,$$

gdzie σ_α jest estymatorem wariancji α-tej współrzędnej zmiennej $\boldsymbol{X}$. Inną możliwością jest tzw. ***reguła Scotta***

$$H_{\alpha\beta} = 0 \,, \quad \sqrt{H_{\alpha\alpha}} = n^{\frac{-1}{m+4}} \sigma_\alpha \,.$$

KDE w praktycznych zastosowaniach często przyspiesza się za pomocą odpowiednich struktur danych do wyszukiwania najbliższych sąsiadów w przestrzeni $\mathbb{R}^m$, tj. zamiast sumować przyczynki od wszystkich punktów $\boldsymbol{x}_i$ dla danego $\boldsymbol{x}$, znajdujemy jego k najbliższych sąsiadów $\boldsymbol{x}$ ze zbioru $\{\boldsymbol{x}_i\}_{i=1}^n$ (np. stosując ***approximate nearest neighbors***) i obliczamy przyczynki do $\hat{p}(\boldsymbol{x})$ tylko od nich.

11 PCA

Analiza składowych głównych (ang. ***Principle Component Analysis***) to liniowa metoda redukcji wymiarowości, która może zostać wykorzystana zarówno w celach wizualzacji wysokowymiarowych danych, jak i do zmniejszenia liczby wymiarów celem ułatwionego ich późniejszego przetwarzania przez inne modele uczenia maszynowego. Zasadnicza idea jest następująca. Mamy macierz $\boldsymbol{X}$ wymiaru $n \times d$, gdzie n to liczba przykładów, a d to liczba cech. Wiemy, że dla każdej macierzy $\boldsymbol{X}$ wymiaru $n \times d$ ($n \geq d$) istnieje jej rozkład SVD

$$\boldsymbol{X} = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^\intercal + \ldots + \sigma_d \boldsymbol{u}_r \boldsymbol{v}_r^\intercal \,,$$

gdzie $r \leq \min\{n, d\}$ to rząd macierzy $\boldsymbol{X}$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0$ – to wartości osobliwe, natomiast$\boldsymbol{u}_i$ to ortonormalne wektory kolumnowe wymiaru n, a $\boldsymbol{v}_i$ to ortonormalne wektory kolumnowe wymiaru d. Możemy w takim razie przybliżyć macierz $\boldsymbol{X}$ za pomocą kilku pierwszych wyrażeń, w szczególności dwóch

$$\boldsymbol{X} \approx \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^\intercal + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^\intercal \,.$$

Zauważmy jednak teraz, iż wektory $\boldsymbol{u}_1$ i $\boldsymbol{u}_2$ możemy potraktować jako kolejne współrzędne 2-D naszego zbioru punktów.

12 Ewaluacja i selekcja modeli

12.1 Przygotowanie danych

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

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

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

12.2 Metryki do oceny regresji i klasyfikacji

Zasadniczo, aby ocenić predykcje modelu używamy odpowiednich metryk, których wartości określają jak dobry jest model.

W przypadku regresji najczęściej używanymi metrykami są RMSE (ang. ***Root Mean Squared Error***) oraz MAE (ang. ***Mean Absolute Error***) zdefiniowane odpowiednio jako

$$\mathrm{RMSE} := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\mathrm{MAE} := \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Metryki te mają jednakową jednostkę jak predykcje. Jeśli chcielibyśmy mieć liczbę względną określającą jakość modelu to mamy do dyspozycji metryki MAPE (ang. ***Mean Absolute Percentage Error***) oraz SMAPE (ang. ***Symmetric Mean Absolute Percentage Error***) zdefiniowane odpowiednio jako

$$\mathrm{MAPE} := \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

$$\mathrm{SMAPE} := \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}$$

Obie te metryki mają zakres od 0 do 1, przy czym niższa wartość oznacza lepszy model. Metryki te mają jednak szereg problemów, z których najpoważniejsze to: problemy, gdy wartości są bliskie 0, asymetryczne traktowanie predykcji za dużych oraz za małych. Z tych powodów znacznie lepszą względną metryką jest MASE (ang. ***Mean Absolute Scaled Error***)

$$\mathrm{MASE} := \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i - \overline{y}|}\,,$$

gdzie $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Metryka MASE jest zatem względnym błędem MAE jaki popełnia nasz model w stosunku do modelu naiwnego, który przewiduje zawsze wartość średnią.

W przypadku zadania klasyfikacji binarnej naszym celem dla danego wektora cech jest zwrócenie jednej z dwóch klas, które będziemy nazywać klasą pozytywną i negatywną. O ile w przypadku regresji pomiar jakości modelu był całkiem prosty, o tyle w przypadku klasyfikacji sytuacja jest nieco bardziej skomplikowana. Zauważmy bowiem, iż mamy 4 możliwości odpowiedzi klasyfikatora

- ***True Positive (TP)*** – poprawnie zaklasyfikowaliśmy klasę pozytywną jako pozytywną
- ***True Negative (TN)*** – poprawnie zaklasyfikowaliśmy klasę negatywną jako negatywną
- ***False Positive (FP)*** – niepoprawnie zaklasyfikowaliśmy klasę negatywną jako pozytywną
- ***False Negative (FN)*** – niepoprawnie zaklasyfikowaliśmy klasę pozytywną jako negatywną

Na podstawie ilości TP, TN, FP i FN w zbiorze testowym możemy wykreślić tzw. ***macierz pomyłek*** (ang. ***confusion matrix***) pokazującą ilość każdej z możliwości. Następnie możemy obliczyć różne stosunki tych wartości, aby uzyskać różne metryki. Najbardziej standardowymi są ***accuracy***, ***precision*** oraz ***recall*** (lub inaczej sensitivity) zdefiniowane jako

$$\begin{aligned}\text{Accuracy} &:= \frac{\text{TP}+\text{TN}}{n}\\ \text{Precision} &:= \frac{\text{TP}}{\text{TP}+\text{FP}}\\ \text{Recall} &:= \frac{\text{TP}}{\text{TP}+\text{FN}}\end{aligned}$$

Wartość accuracy mówi po prostu jaki stosunek przykładów został poprawnie zaklasyfikowany (zauważmy tutaj, że $\text{TP}+\text{TN}+\text{FP}+\text{FN} = n$). Nie jest to jednak dobra miara jakości, gdy nasz zbiór jest niezbalansowany, tj. zawiera więcej przykładów określonej klasy.

Wartość precision określa jak pewny jest klasyfikator przy wykrywaniu klasy pozytywnej, natomiast recall mówi o tym jak dobrze klasyfikator wyławia przykłady pozytywne. Zauważmy jednak, iż nie możemy stosować żadnej z tych metryk w odosobnieniu. Istotnie klasyfikator, który zwraca zawsze klasę pozytywną ma maksymalny recall, a klasyfikator, który zwraca zawsze klasę negatywną ma nieokreślony precision (i jest oczywiście beznadziejnym klasyfikatorem). Musimy więc zawsze ewaluować model na obu tych metrykach i jedynie dobry wynik obu z nich mówi o jakości klasyfikatora. Oczywiście czasami chcielibyśmy określić jakość modelu za pomocą jednej liczby, a niekoniecznie sprawdzać zawsze macierz pomyłek (choć jest to bardzo użyteczne) lub podawać wartości dwóch metryk. Metryką, która łączy precision i recall jest F_β–***score*** zdefiniowany jako

$$F_\beta := (1+\beta^2)\frac{\text{Precision}\cdot\text{Recall}}{\beta^2\cdot\text{Precision}+\text{Recall}},$$

gdzie β określa ile razy bardziej ważny jest recall od precision. Typowo używa się F_1–score

$$F_1 = 2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}.$$

Wiele klasyfikatorów oprócz twardych predykcji zwraca również rozkład prawdopodobieństwa nad klasami. W przypadku klasyfikacji binarnej jest to oczywiście rozkład zero-jedynkowy z parametrem p określającym prawdopodobieństwo klasy pozytywnej dla danego wektora cech. Standardowo oczywiście twardą predykcją jest ta z klas, która ma większe prawdopodobieństwo, czyli (co równoważne) predykcją jest klasa pozytywna jeśli $p > 0.5$. W niektórych problemach chcemy jednak zmienić ten próg i dokonać tzw. ***threshold tuning***. Wykresem, który pozwala na dokonanie tuningu progu jest ***krzywa ROC*** (ang. ***Receiver Operatic Characteristic curve***), która jest krzywą parametryczną wyznaczoną przez punkty $(\text{FPR}(\text{threshold}), \text{TPR}(\text{threshold}))$ dla progów z zakresu $[0; 1]$, gdzie

$$\text{TPR} := \frac{\text{TP}}{\text{TP}+\text{FN}}, \quad \text{FPR} := \frac{\text{FP}}{\text{FP}+\text{TN}}.$$

Metryką niezależną od wybranego progu jest tzw. ***AUROC*** (ang. ***Area under ROC curve***) zdefiniowany jako pole powierzchni pod krzywą ROC dla danego klasyfikatora. Zauważmy, że klasyfikator losowy, który zwraca zawsze klasę pozytywną z prawdopodobieństwem równym wartości progu ma wartość AUROC równą 0.5, natomiast idealny klasyfikator, który niezależnie od wartości progu klasyfikuje wszystkie przykłady poprawnie ma AUROC równy 1.

Inną analogiczną metryką jest ***AUPRC***, gdzie zamiast krzywej ROC stosujemy krzywą ***PRC*** (z ang. ***Precision–Recall Curve***), w której zamiast TPR i FPR używamy odpowiednio Precision i Recall. Metryka AUPRC jest często wykorzystywana w przypadku klasyfikacji ekstremalnie niezbalansowanej, w której mamy bardzo mało ($< 1\%$) klasy pozytywnej.

W przypadku klasyfikacji wieloklasowej używamy zasadniczo takich samych metryk jak w klasyfikacji binarnej, ale wprowadzamy mikro i makro uśrednianie (ang. ***micro/macro-averaging***). Przez

TP_k będziemy rozumieć liczbę prawidłowo zaklasyfikowanych przykładów z klasy k, FP_k to liczba przykładów z innych klas, które zaklasyfikowaliśmy nieprawidłowo jako k–tą klasę, FN_k to liczba przykładów z klasy k, które zaklasyfikowaliśmy jako inną klasę. Wówczas odpowiednie metryki mają postać

$$\text{MicroPrecision} := \frac{\sum_k \mathrm{TP}_k}{\sum_k \mathrm{TP}_k + \sum_k \mathrm{FP}_k},$$
$$\text{MacroPrecision} := \frac{1}{K}\sum_{k=1}^{K} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}$$

oraz

$$\text{MicroRecall} := \frac{\sum_k \mathrm{TP}_k}{\sum_k \mathrm{TP}_k + \sum_k \mathrm{FN}_k},$$
$$\text{MacroRecall} := \frac{1}{K}\sum_{k=1}^{K} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}.$$

W przypadku klasyfikacji wieloklasowej macierz pomyłek jest macierzą wymiaru $K \times K$, gdzie K jest liczbą klas.

12.3 Tuning hiperparametrów i walidacja skrośna

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

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

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

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

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

Walidacja krzyżowa bez jednego (ang. ***Leave-One-Out Cross-Validation***, LOOCV) jest metodą iteracyjną, której kroki wyglądają następująco:

- wybieramy kolejny element zbioru obserwacji $(\boldsymbol{x}_i, y_i)$;
- zbiorem walidacyjnym jest $\{(\boldsymbol{x}_i, y_i)\}$, a zbiorem uczącym reszta obserwacji;
- metodą zbioru walidacyjnego wyliczamy błąd walidacyjny E_i.

Na koniec obliczamy estymatę błędu testowego

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} E_i.$$

Korzyściami z zastosowania LOOCV jest redukcja obciążenia – uczenie modelu korzysta w każdym kroku z $n-1$ obserwacji – w konsekwencji ewentualne przeszacowanie błędu treningowego jest nieznaczne. Dodatkowo metoda jest ***deterministyczna*** – wielokrotne uruchomienie daje ten sam wynik.

Bootstrap to metoda statystyczna stosowana do oszacowania niepewności związanej z danym estymatorem lub metodą uczenia statystycznego. Łatwo stosuje się w sytuacjach, w których nie ma innych – lepiej uzasadnionych teoretycznie – metod oszacowania zmienności. Mamy do dyspozycji zbiór danych $\mathcal{X}$ o wielkości n, przy pomocy którego wyznaczamy estymatę $\hat{\alpha}$ pewnej statystyki α. Zadanie polega na estymowaniu błędu standardowego $\hat{\alpha}$, tzn. $\sigma(\hat{\alpha})$ przy braku możliwości uzyskania nowych danych. Konstruujemy zbiory danych bootstrap $\mathcal{X}_1, \ldots, \mathcal{X}_B$ o wielkości n przez ***losowanie ze zwracaniem*** z $\mathcal{X}$ dla pewnego dużego B i na każdym z $\mathcal{X}_b$ obliczamy estymatę $\hat{\alpha}_b$. Ostatecznie błąd standardowy (tzw. ***błąd typu bootstrap***) ma postać

$$m(\hat{\alpha}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\alpha}_b$$

$$\sigma(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\alpha}_b - m(\hat{\alpha})\right)^2}.$$

12.4 Selekcja cech w modelach liniowych

Po wytrenowaniu modelu liniowego może okazać się, że część predyktorów nie ma istotnego wpływu na zmienną odpowiedzi. Utrzymywanie tych nieistotnych zmiennych nadmiernie zwiększa złożoność modelu, a więc ich usunięcie, tzn. wyzerowanie odpowiednich współczynników, wpływa dodatnio na interpretowalność modelu. Usuwanie nieistotnych predyktorów z modelu regresji wielokrotnej nazywa się selekcją cech lub selekcją zmiennych.

Selekcja podzbioru polega na wyznaczeniu podzbioru predyktorów, które ocenia się jako najsilniej związane z odpowiedzią. Następnie dopasowuje się model liniowy używając metody najmniejszych kwadratów na zredukowanym zbiorze predyktorów. Do metod selekcji podzbioru należą:

- wybór najlepszego podzbioru;
- metody selekcji krokowej.

W metodzie najlepszego podzbioru dopasowujemy model z każdą możliwą kombinacją spośród wszystkich p predyktorów. Spośród otrzymanych 2^p modeli wybieramy najlepszy (najbardziej optymalny według przyjętego kryterium). Zazwyczaj postępuje się według algorytmu:

1. Za $\mathcal{M}_0$ przyjmujemy ***model zerowy*** (bez predyktorów).
2. Dla $k = 1, \ldots, p$:
   (a) dopasowujemy wszystkie $\binom{p}{k}$ modeli z k predyktorami;
   (b) wybieramy $\mathcal{M}_k$, który ma najmniejszy RSS lub największy R^2.
3. Spośród $\mathcal{M}_0, \ldots, \mathcal{M}_p$ wybieramy najbardziej optymalny według przyjętego kryterium.

Ani RSS, ani R^2 nie są dobrymi kryteriami porównywania modeli o różnej liczbie predyktorów, ponieważ – jako miary związane z treningowym MSE – mają lepsze wartości dla modeli o większej liczbie zmiennych. Potrzebne są zatem miary estymujące błąd testowy. Istnieją dwa powszechnie używane podejścia: podejście pośrednie, w którym wprowadzamy poprawki do błędu treningowego uwzględniające potencjalnie nadmierne obciążenie związane z przeuczeniem; podejście bezpośrednie, w którym estymujemy błąd testowy przy pomocy zbioru walidacyjnego albo walidacji krzyżowej. Oczywistą wadą metody wyboru najlepszego podzbioru jest potencjalnie znaczny koszt obliczeniowy. Alternatywą są metody selekcji krokowej, które przeszukują znacznie mniejszy zbiór możliwych modeli.

Algorytm selekcji krokowej w przód ma postać

- $\mathcal{M}_0$ jest modelem zerowym jak poprzednio.
- Dla $k = 0, \ldots, p-1$:
  1. rozważamy wszystkie $p-k$ modeli dodających jeden predyktor do $\mathcal{M}_k$;
  2. wybieramy $\mathcal{M}_{k+1}$, który ma najmniejszy RSS lub największy R^2
- Spośród $\mathcal{M}_0, \ldots, \mathcal{M}_p$ wybieramy najbardziej optymalny względem przyjętego kryterium.

Selekcja krokowa w przód dopasowuje $1 + p(p+1)/2$ modeli (zamiast 2^p), przy czym efektywnie przeszukiwana podprzestrzeń zawiera istotnie więcej modeli. Nie ma gwarancji odnalezienia globalnie najlepszego modelu.

Algorytm selekcji krokowej wstecz ma postać

- $\mathcal{M}_p$ jest modelem pełnym, czyli zawierającym wszystkie predyktory.
- Dla $k = p, p-1, \ldots, 1$:
  1. rozważamy wszystkie k modeli eliminujących jeden predyktor z $\mathcal{M}_k$;
  2. wybieramy $\mathcal{M}_{k-1}$, który ma najmniejszy RSS lub największy R^2.
- Spośród $\mathcal{M}_0, \ldots, \mathcal{M}_p$ wybieramy najbardziej optymalny względem przyjętego kryterium.

W podejściach hybrydowych zmienne są sekwencyjnie dodawane do modelu tak jak w selekcji krokowej w przód. Po każdym dodaniu nowej zmiennej, jedna nieistotna zmienna może zostać usunięta — podobnie jak w selekcji krokowej wstecz. Podejścia te starają się ściślej naśladować wybór najlepszego podzbioru przy zachowaniu niskiego kosztu obliczeniowego metod krokowych.

12.5 Kalibracja modeli