Elementy Statystycznego Uczenia Maszynowego

Regresja Liniowa (kontynuacja)

Regresja grzbietowa

Rozważmy Bayesowską regresję liniową, w której prior ma postać:

$$\mathbf{w} \sim N\left(\mathbf{0}, \tau^2 \mathbf{I}\right)$$

- A więc elementy wektora parametrów są (a priori) wzajemnie niezależne a ich amplitudy są rzędu τ.

Wiarygodność pozostaje bez zmian:

$$\mathbf{y} \mid \mathbf{w} \sim N\left(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}\right)$$

Regresja grzbietowa

Interesuje nas estymata maksimum a posteriori parametrów $\mathbf{w}$.

Ponieważ:

$$p\left(\mathbf{w} \mid D\right) \propto p\left(\mathbf{w} \mid \mathbf{0}, \tau^2\mathbf{I}\right) p\left(\mathbf{y} \mid \mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}\right)$$

to estymatą MAP jest:

$$\mathbf{w}_{\mathrm{MAP}} = \arg\max_{\mathbf{w}} \left[p\left(\mathbf{w} \mid \mathbf{0}, \tau^2\mathbf{I}\right) p\left(\mathbf{y} \mid \mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}\right)\right]$$

Regresja grzbietowa

Korzystając z monotoniczności funkcji logarytmicznej otrzymujemy:

$$\begin{aligned}\mathbf{w}_{\text{MAP}} &= \arg\max_{\mathbf{w}} \left[\log p\left(\mathbf{w} \mid \mathbf{0}, \tau^2\mathbf{I}\right) + \log p\left(\mathbf{y} \mid \mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}\right)\right] \\ &= \arg\max_{\mathbf{w}} \left[-\tau^{-2}\mathbf{w}^{\mathsf{T}}\mathbf{w} - \sigma^{-2}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)^{\mathsf{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right) + const\right] \\ &= \arg\min_{\mathbf{w}} \left[\tau^{-2}\mathbf{w}^{\mathsf{T}}\mathbf{w} + \sigma^{-2}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)^{\mathsf{T}}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)\right]\end{aligned}$$

Wyrażenie w nawiasach można pomnożyć przez σ^2 – nie wpłynie to na położenie minimum.

- Bo σ^2 to dodatnia stała.

Regresja grzbietowa

Estymata MAP w tym modelu przyjmuje więc ostatecznie postać:

$$\mathbf{w}_{\mathrm{MAP}} = \arg\min_{\mathbf{w}} \left[(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}} (\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{\sigma^2}{\tau^2} \mathbf{w}^{\mathsf{T}} \mathbf{w} \right]$$

$$= \arg\min_{\mathbf{w}} \left[(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}} (\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{\sigma^2}{\tau^2} \|\mathbf{w}\|^2 \right]$$

Człon minimalizowany w metodzie najmniejszych kwadratów.

Człon regularyzujący rozwiązanie, który ogranicza amplitudy parametrów (wag).

Regresja grzbietowa

Model ten możemy również zapisać w postaci:

$$\mathbf{w}_{\mathrm{RR}} = \arg\min_{\mathbf{w}} \left[(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathsf{T}} (\mathbf{y} - \mathbf{X}\mathbf{w}) + \gamma \|\mathbf{w}\|^2 \right]$$

gdzie $\gamma > 0$ to stała regularyzująca.

Jest to tak zwana **regresja grzbietowa** (ang. ***ridge regression*** – stąd nazwa estymaty w powyższym wzorze: $\mathbf{w}_{\mathrm{RR}}$).

Regresja grzbietowa

Regresja grzbietowa posiada rozwiązanie w postaci zamkniętej.

- Zgodnie z dotychczasowym wyprowadzeniem, rozwiązaniem jest estymata MAP dla Bayesowskiej regresji liniowej z priorem postaci:

$$\mathbf{w} \sim N\left(\mathbf{0}, \tau^2 \mathbf{I}\right)$$

- Ponieważ prior i wiarygodność są rozkładami normalnymi, posterior również jest rozkładem normalnym.

Regresja grzbietowa

Regresja grzbietowa posiada rozwiązanie w postaci zamkniętej.

- W rozkładzie normalnym punktem o największej gęstości prawdopodobieństwa (modą) jest wartość oczekiwana.
- A więc:

$$\mathbf{w}_{\mathrm{RR}} = \boldsymbol{\mu}_n = \boldsymbol{\Sigma}_n \left[\frac{1}{\sigma^2} \mathbf{X}^\mathsf{T} \mathbf{y} + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right]$$

$$= \left[\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \mathbf{X}^\mathsf{T} \mathbf{X} \right]^{-1} \left[\frac{1}{\sigma^2} \mathbf{X}^\mathsf{T} \mathbf{y} + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right]$$

Regresja grzbietowa

Regresja grzbietowa posiada rozwiązanie w postaci zamkniętej.

- Podstawiając parametry priora otrzymujemy:

$$\begin{aligned}
\mathbf{w}_{\mathrm{RR}} &= \left[\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}\right]^{-1}\left[\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{y} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0\right] \\
&= \left[\frac{1}{\tau^2}\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}\right]^{-1}\left[\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{y} + \frac{1}{\tau^2}\mathbf{I}\mathbf{0}\right] \\
&= \left[\frac{\sigma^2}{\tau^2}\mathbf{I} + \mathbf{X}^\mathsf{T}\mathbf{X}\right]^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}
\end{aligned}$$

Regresja grzbietowa

Regresja grzbietowa posiada rozwiązanie w postaci zamkniętej.

- Rozwiązanie to możemy również zapisać w postaci:

$$\mathbf{w}_{\mathrm{RR}} = \left[\gamma \mathbf{I} + \mathbf{X}^{\mathsf{T}}\mathbf{X}\right]^{-1} \mathbf{X}^{\mathsf{T}}\mathbf{y}$$

- gdzie $\gamma > 0$ to stała (hiper-parametr) określająca siłę regularyzacji.

Wprowadzenie do procesów Gaussowskich

Własności macierzy kowariancji

Przypomnijmy: macierzą kowariancji wielowymiarowej zmiennej losowej $\mathbf{x}$ o wartości oczekiwanej $\boldsymbol{\mu}$ jest macierz:

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

- Własność I: każda macierz kowariancji jest nieujemnie określona, tj. dla każdego wektora $\mathbf{v}$:

$$\mathbf{v}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{v} \geq 0$$

Własności macierzy kowariancji

Dowód:

- Zwróć uwagę, że $\mathbf{v}^\mathsf{T}(\mathbf{x}-\boldsymbol{\mu})$ to skalar. Oznaczmy go przez z.
- Wówczas:

$$\begin{aligned}\mathbf{v}^\mathsf{T}\boldsymbol{\Sigma}\mathbf{v} &= \mathbf{v}^\mathsf{T}\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\right]\mathbf{v} \\ &= \mathbb{E}\left[\mathbf{v}^\mathsf{T}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\mathsf{T}\mathbf{v}\right] \\ &= \mathbb{E}\left[zz^\mathsf{T}\right] = \mathbb{E}\left[z^2\right] \geq 0\end{aligned}$$

Własności macierzy kowariancji

Przypomnijmy: macierzą kowariancji wielowymiarowej zmiennej losowej $\mathbf{x}$ o wartości oczekiwanej $\boldsymbol{\mu}$ jest macierz:

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

- Własność II: każda (symetryczna) macierz nieujemnie określona jest macierzą kowariancji.

Własności macierzy kowariancji

Dowód:

- Niech $\mathbf{A}$ będzie macierzą nieujemnie określoną. Ponieważ jest ona nieujemnie określona to istnieje macierz $\mathbf{L}$ taka, że:

$$\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$$

- Powyższa dekompozycja to dekompozycja Choleskiego macierzy $\mathbf{A}$.
- Weźmy zmienną losową:

$$\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$$

Własności macierzy kowariancji

Dowód:

- Następnie policzmy macierz kowariancji zmiennej $\mathbf{Lz}$:

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{Lz}} &= \mathbb{E}\left[(\mathbf{Lz}-\boldsymbol{\mu}_{\mathbf{Lz}})(\mathbf{Lz}-\boldsymbol{\mu}_{\mathbf{Lz}})^{\mathsf{T}}\right] \\ &= \mathbb{E}\left[(\mathbf{Lz}-\mathbf{L0})(\mathbf{Lz}-\mathbf{L0})^{\mathsf{T}}\right] \\ &= \mathbb{E}\left[\mathbf{Lzz}^{\mathsf{T}}\mathbf{L}^{\mathsf{T}}\right] = \mathbf{L}\mathbb{E}\left[\mathbf{zz}^{\mathsf{T}}\right]\mathbf{L}^{\mathsf{T}} \\ &= \mathbf{LIL}^{\mathsf{T}} = \mathbf{A}\end{aligned}$$

- $\mathbf{A}$ jest więc macierzą kowariancji zmiennej $\mathbf{Lz}$.

Własności macierzy kowariancji

Przypomnijmy: macierzą kowariancji wielowymiarowej zmiennej losowej $\mathbf{x}$ o wartości oczekiwanej $\boldsymbol{\mu}$ jest macierz:

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

- Własności I i II wskazują, że macierze kowariancji można w pewnym sensie utożsamiać z macierzami nieujemnie określonymi.

Funkcje kowariancji

Niech $k : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ będzie funkcją taką, że $\forall n \in \mathbb{N},\ \forall \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ macierz:

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

jest dodatnio określona.

Wówczas k nazywamy **jądrem dodatnio określonym**.

Funkcje kowariancji

Wówczas k nazywamy **jądrem dodatnio określonym**.

- Po angielsku: ***positive definite kernel***.
- Alternatywnie: jądro Mercera (ang. *Mercer kernel*).

Jądra dodatnio określone można traktować jako „przepisy” na macierz kowariancji.

- Stąd inna popularna nazwa: **funkcje kowariancji**.

Funkcje kowariancji

Warunki jakie musi spełniać k by być funkcją kowariancji podaje twierdzenie Mercera.

Nie będziemy go tu przytaczać, lecz podamy kilka przykładów funkcji kowariancji:

- Jądro Gaussowskie:

$$k\left(\mathbf{x}_1, \mathbf{x}_2\right) = \exp\left[-\frac{\left\|\mathbf{x}_1 - \mathbf{x}_2\right\|^2}{2l^2}\right]$$

- Jądro periodyczne:

$$k\left(\mathbf{x}_1, \mathbf{x}_2\right) = \exp\left[-\frac{2}{l^2}\sin^2\left(\pi\frac{\left|\mathbf{x}_1 - \mathbf{x}_2\right|}{p}\right)\right]$$

Funkcje kowariancji

Nie będziemy go tu przytaczać, lecz podamy kilka przykładów funkcji kowariancji:

- Złożenie funkcji kowariancji z wielomianem o nieujemnych współczynnikach:

$$k' = q \circ k$$

- gdzie k to funkcja kowariancji zaś q to wielomian o nieujemnych współczynnikach.
- Suma lub iloczyn dwóch funkcji kowariancji również jest funkcją kowariancji.

Funkcje kowariancji

Mówiąc o wartościach funkcji kowariancji wygodnie jest posługiwać się uproszczoną notacją.

- Dla dwóch zbiorów wektorów:

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n\} \subset \mathbb{R}^d$$

$$Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_m\} \subset \mathbb{R}^d$$

- będziemy pisać:

$$\mathbf{K}_{XY} = k(X, Y) = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{y}_1) & k(\mathbf{x}_1, \mathbf{y}_2) & \cdots & k(\mathbf{x}_1, \mathbf{y}_m) \\ k(\mathbf{x}_2, \mathbf{y}_1) & k(\mathbf{x}_2, \mathbf{y}_2) & \cdots & k(\mathbf{x}_2, \mathbf{y}_m) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_n, \mathbf{y}_1) & k(\mathbf{x}_n, \mathbf{y}_2) & \cdots & k(\mathbf{x}_n, \mathbf{y}_m) \end{bmatrix}$$

Proces Gaussowski

Niech

$$GP = \left\{ f_{\mathbf{x}};\ \mathbf{x} \in \mathbb{R}^d \right\}$$

będzie rodziną (zbiorem) zmiennych losowych indeksowanych przez punkty z $\mathbb{R}^d$.

Mówimy, że GP jest **procesem Gaussowskim**, jeśli każdy jego skończony podzbiór ma łącznie (wielowymiarowy) rozkład normalny.

Proces Gaussowski

Mówimy, że GP jest **procesem Gaussowskim**, jeśli każdy jego skończony podzbiór ma łącznie (wielowymiarowy) rozkład normalny:

$$\forall n \in \mathbb{N},\ X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d :$$

$$\begin{bmatrix} f_{\mathbf{x}_1} \\ f_{\mathbf{x}_2} \\ \vdots \\ f_{\mathbf{x}_n} \end{bmatrix} \sim N\left(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X\right)$$

Proces Gaussowski

$$\begin{bmatrix} f_{\mathbf{x}_1} \\ f_{\mathbf{x}_2} \\ \vdots \\ f_{\mathbf{x}_n} \end{bmatrix} \sim N\left(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X\right)$$

Zwróć uwagę, że wartość oczekiwana i macierz kowariancji zależą od podzbioru, który wybieramy z procesu.

- Proces Gaussowski możemy więc jednoznacznie zdefiniować podając „przepis” na wartości oczekiwane i macierze kowariancji.

Proces Gaussowski

W praktyce często nie modelujemy wartości oczekiwanej i przyjmujemy: $\boldsymbol{\mu}_X = \mathbf{0}$.

„Przepisem” na macierz kowariancji może być natomiast funkcja kowariancji:

$$\boldsymbol{\Sigma}_X = k(X, X) = \mathbf{K}_{XX}$$

Dobrze, ale co daje nam taka konstrukcja?

Proces Gaussowski

Proces Guassowski daje nam w praktyce **rozkład prawdopodobieństwa nad funkcjami**:

$$f \sim GP\left(\mathbf{0}, k\right)$$

gdzie:

$$f : \mathbb{R}^d \mapsto \mathbb{R}$$

Oczywiście ten rozkład nie „przedstawia” funkcji w postaci analitycznej!

Proces Gaussowski

Jednak dla każdego skończonego zbioru punktów $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n\} \subset \mathbb{R}^d$ proces Gaussowski daje nam rozkład prawdopodobieństwa nad wartościami f w punktach z X:

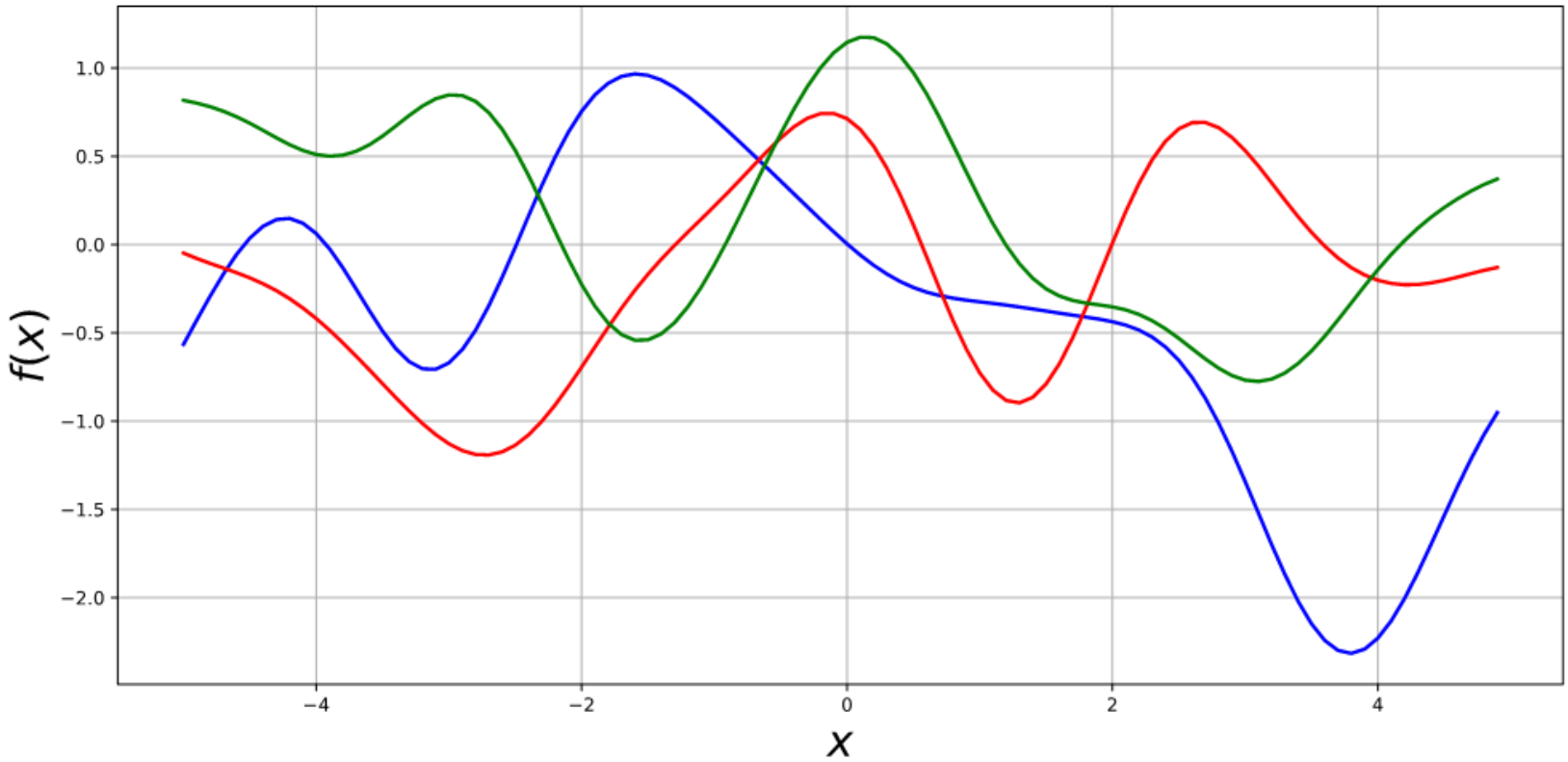
$$\mathbf{f}_X = \begin{bmatrix} f(\mathbf{x}_1) \\ f(\mathbf{x}_2) \\ \vdots \\ f(\mathbf{x}_n) \end{bmatrix} \sim N(\mathbf{0}, \mathbf{K}_{XX})$$

Proces Gaussowski

Jaką rolę pełni tu funkcja kowariancji?

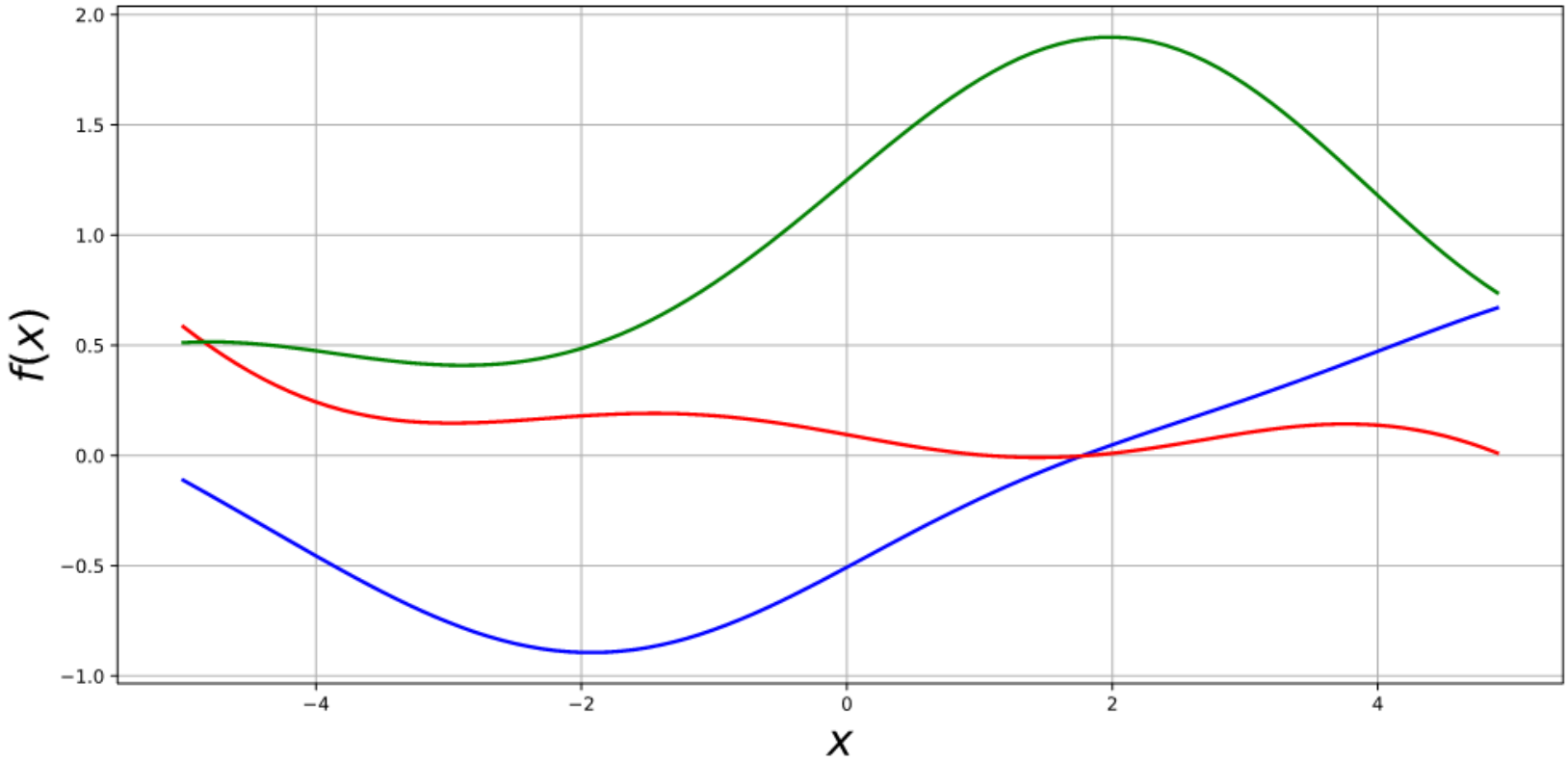
- Funkcja kowariancji w pewnym sensie wyraża zależności pomiędzy wartościami f w różnych punktach $\mathbb{R}^d$.
- Postulując funkcję kowariancji określamy więc jakich cech oczekujemy od f:
  - Periodyczność, gładkość, etc.
- W efekcie uzyskujemy ekspresywny prior nad funkcjami.

Proces Gaussowski



$$k\left(\mathbf{x}_1, \mathbf{x}_2\right) = \exp\left[-\frac{\left\|\mathbf{x}_1 - \mathbf{x}_2\right\|^2}{2l^2}\right], \quad l = 1$$

Proces Gaussowski



$$k\left(\mathbf{x}_1, \mathbf{x}_2\right) = \exp\left[-\frac{\left\|\mathbf{x}_1 - \mathbf{x}_2\right\|^2}{2l^2}\right], \quad l = 3$$

Procesy Gaussowskie w regresji

Procesy Gaussowskie możemy wykorzystać jako prior w Bayesowskim modelowaniu funkcji.

- Na przykład w zagadnieniu regresji – mówimy wówczas o **regresji procesem Gaussowskim** (ang. ***Gaussian Process Regression***).

Procesy Gaussowskie w regresji

Załóżmy, że obserwujemy wartości pewnej nieznanej funkcji f w punktach

$$U = \{\mathbf{u}_1, \mathbf{u}_2, \ldots \mathbf{u}_n\} \subset \mathbb{R}^d$$

przy czym każda obserwowana wartość obarczona jest pewnym błędem:

$$y_i = f(\mathbf{u}_i) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Interesuje nas rozkład prawdopodobieństwa nad wartościami f w punktach:

$$V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_m\} \subset \mathbb{R}^d$$

Procesy Gaussowskie w regresji

Formalnie mamy model postaci:

$$f \sim GP\left(\mathbf{0}, k\right)$$

$$\mathbf{y} \mid f, U \sim N\left(\mathbf{f}_U, \sigma^2 \mathbf{I}\right)$$

Ponieważ błąd pomiaru jest niezależny od obserwacji, to:

$$\operatorname{cov}\left[\mathbf{y} \mid f, U\right] = \operatorname{cov}\left[\mathbf{f}_U\right] + \operatorname{cov}\left[\boldsymbol{\epsilon}\right] = \mathbf{K}_{UU} + \sigma^2 \mathbf{I}$$

Procesy Gaussowskie w regresji

Dalej, zgodnie z definicją procesu Gaussowskiego wartości f nad dowolnym skończonym zbiorem punktów mają rozkład normalny.

Skoro tak, to $\mathbf{y}$ oraz $\mathbf{f}_V$ mają łącznie rozkład normalny:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_V \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{K_{UU}} + \sigma^2\mathbf{I} & \mathbf{K}_{UV} \\ \mathbf{K_{VU}} & \mathbf{K}_{VV} \end{bmatrix}\right)$$

Procesy Gaussowskie w regresji

Z wykładu II wiemy więc, że rozkład warunkowy $p(\mathbf{f}_V \mid \mathbf{y})$ również jest rozkładem normalnym:

$$\mathbf{f}_V \mid \mathbf{y}, U \sim N(\boldsymbol{\mu}_V, \boldsymbol{\Sigma}_V)$$

gdzie (po podstawieniu wartości do wzorów z wykładu II):

$$\boldsymbol{\mu}_V = \mathbf{K}_{VU}\left(\mathbf{K_{UU}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}$$

$$\boldsymbol{\Sigma}_V = \mathbf{K}_{VV} - \mathbf{K}_{VU}\left(\mathbf{K_{UU}} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{K}_{UV}$$

Procesy Gaussowskie w regresji

Rozkład warunkowy $p(\mathbf{f}_V \mid \mathbf{y})$ jest w tym modelu rozkładem predykcyjnym:

- Pozwala nam wnioskować o wartościach funkcji f w punktach ze zbioru V gdy znamy przybliżane wartości f w punktach ze zbioru U.
- Funkcja kowariancji wyraża natomiast nasze założenia a priori co do charakteru funkcji f.