VOLGOGRAD STATE TECHNICAL UNIVERSITY

Машинное обучение и нейросетевые модели

Лекция 8. Вероятностные модели и их применение

Лектор: Кравченя Павел Дмитриевич

Волгоград 2025

VOLGOGRAD STATE TECHNICAL UNIVERSITY

План лекции

1. Недостатки линейных моделей. Переход к обобщенным линейным моделям.
2. Функция связи. Экспоненциальный класс распределений.
3. Лемма о градиенте кумулятивной функции.
4. Подкласс экспоненциальных распределений. Каноническая функция связи.
5. Байесовские линейная, логистическая и пуассоновская регрессии.
6. Понятие моделей смесей. Структура моделей смесей.
7. Смеси гауссовских и категориальных моделей. LDA.
8. Байесовские нейронные сети. Стохастические нейронные сети.
9. Непараметрические модели. Гауссовы процессы. Байесовская оптимизация.
10. Реализация байесовской нейронной сети во фреймворке Pyro.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Недостатки линейных моделей

- Одним из *широко используемых* классов моделей являются *линейные*, в которых *матожидание* целевой переменной определяется *линейной комбинацией признаков*:

$$\mathbb{E}(y \mid \mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle.$$

- Однако, у *линейных моделей* есть существенные *недостатки*:
  1. Линейные модели *не позволяют* учесть нелинейную связь между признаками и целевой переменной.
  2. Дисперсия линейных моделей, как правило, является *константной* и *не зависящей* от признаков.
  3. Линейные модели *не позволяют* обрабатывать дискретную целевую переменную.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Обобщенные линейные модели (Generalized Linear Models)

- Попробуем обобщить *линейные модели*, введя некоторую монотонную функцию $g(t)$, связывающую *матожидание* целевой переменной и *линейную комбинацию* признаков в модели:

$$g\big(\mathbb{E}(y \mid \mathbf{x})\big) = \langle \mathbf{w}, \mathbf{x} \rangle.$$

- Данная функция $g(t)$ называется *функцией связи* (*link function*).
- Таким образом, функция связи позволяет *линеаризовать* матожидание целевой переменной по признакам.
- Таким образом, для определения *конкретной модели* требуются:
  - *Параметризованное* семейство распределений: $p_\theta(y \mid \langle \mathbf{x}, \mathbf{w} \rangle)$.
  - *Функцию связи* $g(t)$.
- Определенная таким образом модель называется обобщенной линейной.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Экспоненциальный класс распределений

- При построении обобщенной линейной модели можно выбрать *любой класс распределений* $p_\theta(y \mid \langle \mathbf{x}, \mathbf{w} \rangle)$ и *любую монотонную функцию связи* $g(t)$, получив при этом *некоторую вероятностную модель*.
- Однако, часто предполагают, что *класс распределений* принадлежит одному из достаточно простых семейств *экспоненциального класса*, который в общем виде выражается следующим образом:

$$p(y \mid \boldsymbol{\theta}) = \frac{1}{h(\boldsymbol{\theta})} f(y) \cdot \exp(\boldsymbol{\theta}^T \mathbf{u}(y)).$$

- Здесь θ – вектор *вещественнозначных* параметров (значениями которых различаются распределения из семейства), $h > 0$, $f > 0$, $\mathbf{u}$ – некоторая *вектор-функция*.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Лемма о градиенте кумулятивной функции

- Для экспоненциального семейства:

$$p(y \mid \boldsymbol{\theta}) = \frac{1}{h(\boldsymbol{\theta})} f(y) \cdot \exp(\boldsymbol{\theta}^T \mathbf{u}(y)) .$$

- Интегрируя обе части равенства по y и *применяя условие нормировки*, получаем:

$$h(\boldsymbol{\theta}) = \int f(y) \cdot \exp\left(\boldsymbol{\theta}^T \mathbf{u}(y)\right) dy .$$

- Вычислим градиент от логарифма $h(\boldsymbol{\theta})$:

$$\nabla_{\boldsymbol{\theta}} \log h(\boldsymbol{\theta}) = \frac{\nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta})}{h(\boldsymbol{\theta})} = \frac{\nabla_{\boldsymbol{\theta}} \int f(y) \cdot \exp\left(\boldsymbol{\theta}^T \mathbf{u}(y)\right) dy}{h(\boldsymbol{\theta})} = \frac{\int f(y) \nabla_{\boldsymbol{\theta}} \exp\left(\boldsymbol{\theta}^T \mathbf{u}(y)\right) dy}{h(\boldsymbol{\theta})} =$$

$$= \frac{\int f(y) \mathbf{u}(y) \exp\left(\boldsymbol{\theta}^T \mathbf{u}(y)\right) dy}{h(\boldsymbol{\theta})} = \int \mathbf{u}(y) \cdot \frac{1}{h(\boldsymbol{\theta})} f(y) \exp\left(\boldsymbol{\theta}^T \mathbf{u}(y)\right) dy = \mathbb{E}_{p(y|\boldsymbol{\theta})}[\mathbf{u}(y)].$$

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Подкласс экспоненциальных распределений для GLM

- В контексте *обобщенных линейных моделей* обычно рассматривают *подкласс* экспоненциального класса, состоящий из семейств, представимых в виде:

$$p(y \mid \theta, \phi) = \exp\left(\frac{y\theta - a(\theta)}{\phi} + b(y, \phi)\right).$$

где θ и ϕ – скалярные параметры, причем ϕ – фиксированный параметр, а значения θ параметризуют распределения из семейства. Величины $a(\theta)$ и $b(y, \phi)$ – некоторые скалярные функции.

- Данное выражение можно переписать в виде, *больше похожим* на канонический вид *экспоненциального класса распределений*:

$$p(y \mid \theta, \phi) = \frac{1}{\exp\left(\frac{a(\theta)}{\phi}\right)} \exp[b(y, \phi)] \cdot \exp\left(\frac{y\theta}{\phi}\right).$$

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Способ определения функции связи

- Поскольку ϕ является фиксированным параметром (константой), имеем:

$$p(y \mid \theta, \phi) = \frac{1}{h(\theta)} f(y) \cdot \exp\left(\boldsymbol{\theta}^T \mathbf{u}(y)\right), \qquad h(\theta) = \exp\left(\frac{a(\theta)}{\phi}\right),$$

- Видно, что функция $\mathbf{u}(y)$ содержит только одну компоненту $u_1(y)$, причём:

$$u_1(y) = \frac{y}{\phi}.$$

- В соответствии с доказанной леммой получаем:

$$\mathbb{E}[u_1(y)] = \nabla_\theta \log h(\theta),$$

$$\frac{\mathbb{E}[y]}{\phi} = \frac{\partial}{\partial \theta}\left(\frac{a(\theta)}{\phi}\right) = \frac{1}{\phi}\frac{\partial}{\partial \theta}\big(a(\theta)\big).$$

$$\mathbb{E}[y] = \frac{\partial}{\partial \theta}\big(a(\theta)\big) = a'(\theta).$$

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Каноническая функция связи

- Как будет выглядеть выражение для $p(y \mid \mathbf{x})$?
- Единственное, что может *зависеть* от $\mathbf{x}$ в выражении *экспоненциального класса*, – это параметр θ. Положим $\theta = \langle \mathbf{w}, \mathbf{x} \rangle$. Тогда:

$$\mathbb{E}[y \mid \mathbf{x}] = a'(\langle \mathbf{w}, \mathbf{x} \rangle).$$

- Но для функции связи справедливо соотношение:

$$g\big(\mathbb{E}(y \mid \mathbf{x})\big) = \langle \mathbf{w}, \mathbf{x} \rangle.$$

- Отсюда можно *однозначно* определить *функцию связи*:

$$g = (a')^{-1}.$$

- *Функция связи*, введенная таким образом, называется *канонической*.
- Поэтому, чтобы определить *функцию связи*, нужно вычислить $a'(\langle \mathbf{w}, \mathbf{x} \rangle)$.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Пример: байесовская линейная регрессия

- Рассмотрим *байесовскую линейную регрессию*:

$$p(y \mid \mathbf{x}, \mathbf{w}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left[-\frac{(y - \langle \mathbf{w}, \mathbf{x} \rangle)^2}{2\sigma^2}\right].$$

- Вычислим функцию связи для данного семейства распределений. Оно принадлежит к экспоненциальному классу. Попробуем представить его в виде:

$$p(y \mid \mu, \sigma^2) = \exp\left(\frac{y\mu - a(\mu)}{\sigma^2} + b(y, \sigma^2)\right).$$

$$p(y \mid \mu, \sigma^2) = \exp\left[-\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{(y-\mu)^2}{2\sigma^2}\right] = \exp\left[-\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right] =$$

$$= \exp\left[\frac{y\mu - (\mu^2/2)}{\sigma^2} - \left(\log\frac{1}{\sqrt{2\pi}\sigma} + \frac{y^2}{2\sigma^2}\right)\right] \Rightarrow \; a(\mu) = \frac{\mu^2}{2}.$$

- Тогда $a'(\mu) = \mu, \; \mathbb{E}(y \mid \mathbf{w}, \mathbf{x}) = g^{-1}(\langle \mathbf{w}, \mathbf{x} \rangle) = a'(\langle \mathbf{w}, \mathbf{x} \rangle) = \langle \mathbf{w}, \mathbf{x} \rangle.$

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Пример: байесовская логистическая регрессия

- Пусть $\mu = \mathbb{E}[\langle \mathbf{x}, \mathbf{w} \rangle]$. Рассмотрим *байесовскую* *логистическую регрессию*:

$$p(y \mid \mu) = \mu^y (1-\mu)^{1-y}.$$

- Здесь используется *семейство распределений Бернулли*, которое тоже принадлежит к *экспоненциальному классу*.

$$p(y \mid \mu) = \mu^y (1-\mu)^{1-y} = \exp[\log(\mu^y \cdot (1-\mu)^{1-y})] = \exp\left[\log\left(\frac{\mu^y}{(1-\mu)^{y-1}}\right)\right] =$$

$$= \exp\left[\log\left(\left(\frac{\mu}{1-\mu}\right)^y (1-\mu)\right)\right] = \exp\left[y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)\right] = \exp\left[\frac{y \log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)}{1} + 0\right]$$

- Видно, что $\theta = \log\left(\frac{\mu}{1-\mu}\right)$, $a(\theta) = -\log(1-\mu) \implies \mu = \frac{1}{1+e^{-\theta}}$, $a(\theta) = \log(1+e^{\theta})$.
- Тогда: $a'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = \frac{1}{1+e^{-\theta}} = \sigma(\theta)$, $\mathbb{E}(y \mid \mathbf{w}, \mathbf{x}) = g^{-1}(\langle \mathbf{w}, \mathbf{x} \rangle) = a'(\langle \mathbf{w}, \mathbf{x} \rangle) = \sigma(\langle \mathbf{w}, \mathbf{x} \rangle)$.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Пример: байесовская пуассоновская регрессия

- Пусть $\mu = \mathbb{E}[\langle \mathbf{x}, \mathbf{w} \rangle]$. Рассмотрим *байесовскую пуассоновскую регрессию*:

$$p(y \mid \mu) = \frac{e^{-\mu}\mu^y}{y!}, \qquad y \in \mathbb{N}.$$

- Здесь используется *семейство распределений Пуассона*, которое тоже принадлежит к *экспоненциальному классу*.

$$p(y \mid \mu) = \frac{e^{-\mu}\mu^y}{y!} = \exp[-\mu + y\log\mu - \log y!] = \exp\left[\frac{y\log\mu - \mu}{1} - \log y!\right].$$

- Видно, что $\theta = \log\mu$, $a(\theta) = \mu \implies \mu = e^\theta,\ a(\theta) = e^\theta$.
- Тогда: $a'(\theta) = e^\theta,\ \mathbb{E}(y \mid \mathbf{w}, \mathbf{x}) = g^{-1}(\langle \mathbf{w}, \mathbf{x} \rangle) = a'(\langle \mathbf{w}, \mathbf{x} \rangle) = e^{\langle \mathbf{w}, \mathbf{x} \rangle}$.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Понятие моделей смесей (Mixture models)

- *Моделью смесей* называют *вероятностную модель* для представления подгрупп объектов в общей группе без требований, чтобы информация о *принадлежности* к конкретной подгруппе *явно* содержалась в признаковом описании объекта.
- *Модель* соответствует *смеси распределений*, которое представляет собой распределение *вероятностей наблюдений* в общей совокупности.
- Часто *модели смесей* используются для описания *мультимодальных данных* (содержащих несколько регионов с высокой плотностью вероятности).
- Модели смесей используются для *кластеризации* (так называемая кластеризация на основе модели), а также для *оценки плотности*.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Структура моделей смесей

- *Модель смеси* обычно состоит из следующих *компонентов*:
  - N наблюдаемых случайных переменных, каждая из которых имеет распределение смеси K компонент (все компоненты принадлежат к *одному параметрическому семейству*, но различаются *параметрами*).
  - N латентных случайных переменных, определяющих *компоненту из смеси* для каждого *наблюдения*. Каждая переменная распределена по *K-мерному категориальному распределению*.
  - Множество K весов смеси, представляющих собой *вероятности*, сумма которых равна единице.
  - Множество K параметров, каждый из которых определяет параметр соответствующего компонента смеси.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Структура моделей смесей

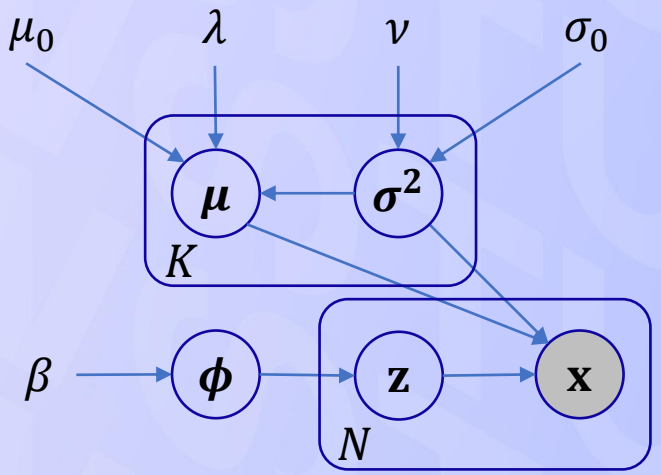
- α – совместный гиперпараметр для *параметров компонент смеси*.
- β – совместный гиперпараметр для *весов смеси*.
- $F(\mathbf{x} \mid \boldsymbol{\theta})$ – распределение вероятностей *наблюдаемых переменных*.
- $H(\boldsymbol{\theta} \mid \alpha)$ – *априорное* распределение *параметров компонент смеси*.
- $\theta_{i=1..K} \sim H(\boldsymbol{\theta} \mid \alpha)$ – параметр распределения наблюдаемой переменной, связанной с i-той компонентой смеси.
- $\phi_{i=1..K}$ – вес i-той компоненты смеси, $\sum_i \phi_i = 1$.
- $z_{i=1..N}$ – компонента, соответствующая i-той наблюдаемой переменной.
- $x_{i=1..N} \sim F(\boldsymbol{\theta}_{z_i})$ – i-тая наблюдаемая переменная.

Распределение $F(\mathbf{x} \mid \boldsymbol{\theta})$ часто является *нормальным* или *категориальным*.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Пример: смесь гауссовских моделей

- $\theta_{i=1..K} = \{\mu_{i=1..K}, \sigma^2_{i=1..K}\}$;
- $\mu_{i=1..K} \sim \mathcal{N}(\mu_0, \lambda\sigma_i^2)$;
- $\sigma^2_{i=1..K} \sim \text{Inv}\mathcal{G}(\nu, \sigma_0^2)$;
- $\boldsymbol{\phi} \sim \text{Symm}\mathcal{D}ir_K(\beta)$;
- $z_{i=1..N} \sim \mathcal{C}at(\boldsymbol{\phi})$;
- $x_{i=1..N} \sim \mathcal{N}(\mu_{z_i}, \sigma^2_{z_i})$.

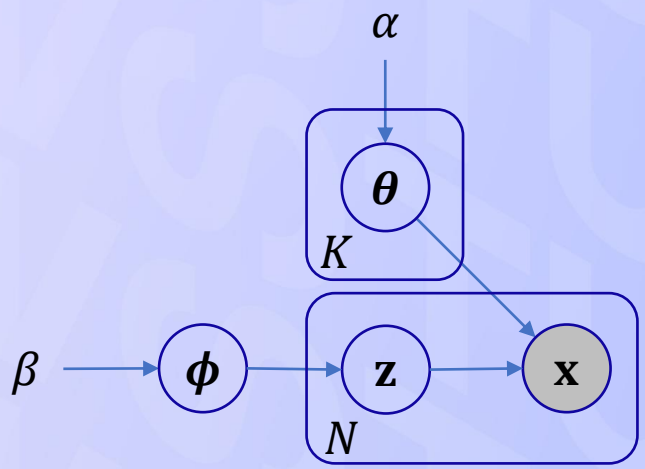


- *Вероятностный процесс* устроен следующим образом. Сначала находится семпл весов компонент смеси $\boldsymbol{\phi}$. Затем семплируется компонента смеси $\mathbf{z}$, и определяются параметры нормального распределения μ и σ^2, которые соответствуют данной компоненте. Затем формируется значение $\mathbf{x}$.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Пример: смесь категориальных моделей

- V – размерность категориальных наблюдаемых переменных;
- $\boldsymbol{\theta}_{i=1..K} = \mathrm{Symm}\mathcal{D}ir_V(\alpha);\ \boldsymbol{\theta}_i \in \mathbb{R}^V;\ \sum_j \theta_{ij} = 1;$
- $\boldsymbol{\phi} \sim \mathrm{Symm}\mathcal{D}ir_K(\beta);$
- $z_{i=1..N} \sim \mathcal{C}at(\boldsymbol{\phi});$
- $x_{i=1..N} \sim \mathcal{C}at(\boldsymbol{\theta}_{z_i});$



- *Вероятностный процесс* устроен следующим образом. Сначала находится семпл весов компонент смеси $\boldsymbol{\phi}$. Затем семплируется компонента смеси $\boldsymbol{z}$, и определяются параметры категориального распределения $\boldsymbol{\theta}$, которые соответствуют данной компоненте. Затем формируется значение $\mathbf{x}$.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

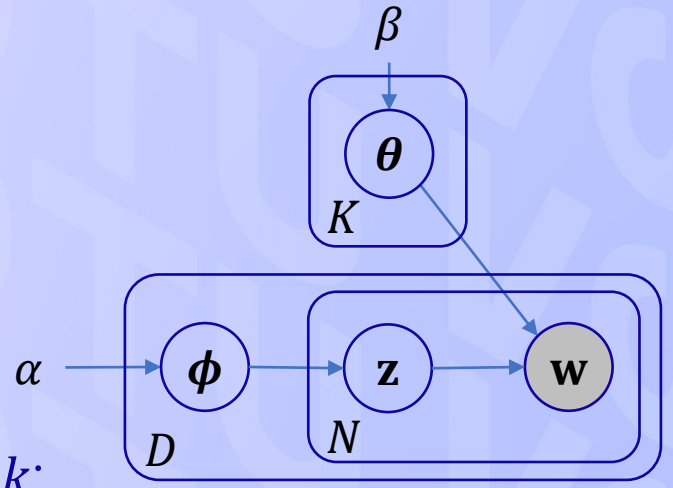
Пример: Латентное размещение Дирихле (LDA)

- Модель латентного размещения Дирихле (LDA) относится к тематическим моделям, которые предназначены для описания текстов с точки зрения их *тематик*. LDA — это вероятностная модель *порождения* текста.
- Набор текстов (корпус) состоит из D *документов*. Каждый документ включает N *слов*.
- При этом *порядок употребления* терминов в документе игнорируется (модель «мешка слов»).
- Основное предположение тематической модели LDA состоит в том, что *каждый документ* имеет несколько тематик, смешанных в некоторой пропорции. Всего имеется K тематик. Тематика рассматривается как распределение вероятностей в пространстве слов из общего словаря.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Пример: Латентное размещение Дирихле (LDA)

- $w \in \{1,2,\dots,W\}$ – номер слова в словаре;
- $k \in \{1,2,\dots,K\}$ – номер тематики;
- $d \in \{1,2,\dots,D\}$ – номер документа в корпусе;
- $\mathbf{w}_d = [w_{d,1}, w_{d,2}, \dots, w_{d,N}]$ – слова в документе d;
- $\mathbf{z}_d = [z_{d,1}, z_{d,2}, \dots, z_{d,N}]$ – тематики слов в документе d;
- $\boldsymbol{\theta}_k = [\theta_{k,1}, \theta_{k,2}, \dots, \theta_{k,W}]$ – вероятности слов в тематике k;
- $\boldsymbol{\phi}_d = [\phi_{d,1}, \phi_{d,2}, \dots, \phi_{d,K}]$ – вероятности тематик в документе d;
- $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \dots, \boldsymbol{\phi}_D]^K \in \mathbb{R}^{D \times K}$ – вероятности тематик во всех документах;
- $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \dots, \boldsymbol{\Theta}_K]^W \in \mathbb{R}^{K \times W}$ – вероятности слов во всех тематиках.



VOLGOGRAD STATE TECHNICAL UNIVERSITY

Другие возможные распределения компонент смеси

- Возможные *распределения компонент смеси:*
  - *Биномиальное распределение* количества «положительных событий» (*например*, успехов, голосов «да» и т. д.) при фиксированном общем количестве вхождений.
  - *Распределение Пуассона* для количества *повторений* события за *заданный период времени* (для события, которое характеризуется фиксированной частотой возникновения).
  - Экспоненциальное распределение характеризует *время* до появления *следующего события* (для события, которое характеризуется фиксированной частотой возникновения).

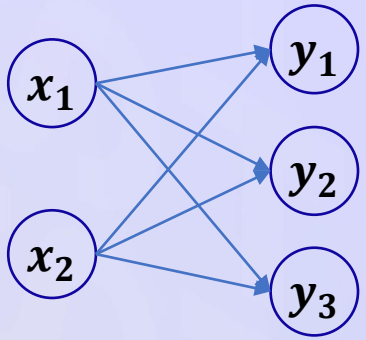
VOLGOGRAD STATE TECHNICAL UNIVERSITY

Другие возможные распределения компонент смеси

- Возможные *распределения* *компонент смеси:*
  - *Логарифмически нормальное* распределение для *положительных действительных* чисел, которые, как предполагается, растут экспоненциально (например, доходы или цены).
  - *Многомерное нормальное распределение* (также известное как многомерное распределение Гаусса) для векторов коррелирующих событий, которые *индивидуально* распределены по Гауссу.
  - *Многомерное t-распределение Стьюдента* для векторов коррелированных результатов с тяжелым хвостом, и т.д.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

- Искусственная нейронная сеть, в силу универсальной теоремы аппроксимации, позволяет воспроизвести произвольную функцию $\mathbf{y} = \Phi(\mathbf{x})$.



$$y_1 = \Phi(w_{11}x_1 + w_{12}x_2 + w_{10})$$

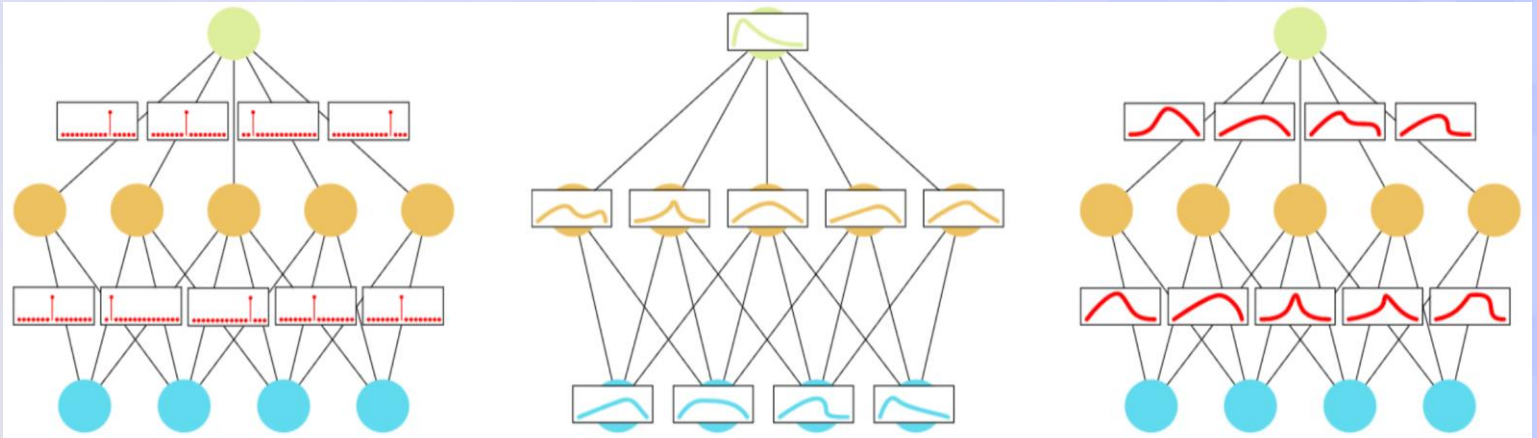
$$y_2 = \Phi(w_{21}x_1 + w_{22}x_2 + w_{20})$$

$$\mathbf{y} = \Phi(\mathbf{W}\mathbf{x})$$

- В простейшей архитектуре сетей прямого распространения каждый слой представлен как *линейное преобразование*, за которым следует нелинейная операция Φ, также известная как *функция активации*.
- Байесовские нейронные сети (BNN) представляют собой *стохастические* нейронные сети, обученные с использованием байесовского подхода.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

- Стохастические нейронные сети — это тип нейросетей, построенный путем введения в сеть стохастических компонентов. Он реализуется путём введения либо *стохастических активаций*, либо *стохастических весов* для моделирования *нескольких* возможных моделей с соответствующим распределением вероятностей. Таким образом, BNN можно рассматривать как частный случай *ансамблевого обучения*.



VOLGOGRAD STATE TECHNICAL UNIVERSITY

Построение байесовской нейронной сети

- Первым шагом при разработке байесовской нейронной сети является выбор *архитектуры* глубокой сети, то есть *функциональной модели*:

$$\mathbf{y} = \Phi_{\boldsymbol{w}}(\mathbf{x}).$$

- Затем необходимо выбрать *стохастическую модель*, то есть *априорное распределение* возможных *параметров* модели и *априорную уверенность* в предсказательной силе модели:

$$p(\mathbf{w}), \qquad p(\mathbf{y} \mid \mathbf{x}, \mathbf{w}).$$

- Выбор *стохастической* модели *байесовской* нейронной сети некоторым образом *эквивалентен выбору функции потерь* при обучении нейронной сети с *точечной оценкой*.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Виды неопределенности в байесовском моделировании

- В байесовском моделировании выделяют два основных вида неопределенности:
  - *Алеаторическая неопределенность* отражает шум, присущий наблюдениям. Это может быть, например, шум датчиков или шум движения. Данную неопределенность *невозможно уменьшить*, даже если собрать больше данных.
  - *Эпистемическая неопределенность* объясняет неопределенность в *параметрах модели*, которая отражает наше незнание того, как модель сгенерировала собранные данные. Эту неопределенность можно объяснить, имея достаточно данных, и ее часто называют неопределенностью модели.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Реализация байесовской нейронной сети во фреймворке Pyro

- Для работы с *байесовской нейросетью* в Pyro нужно воспользоваться классом PyroModule:

- class Linear(nn.Module):
+ class Linear(PyroModule):
  def __init__(self, in, out):
  super().__init__()
  self.weight = ...
  self.bias = ...
  ...
	- linear = Linear(5, 2)
	+ linear = PyroModule[Linear](5, 2)
		linear = Linear(5, 2)
		assert isinstance(linear, nn.Module)
		assert not isinstance(linear, PyroModule)
		
		to_pyro_module_(linear)

- Далее нужно заменить атрибуты *nn.Parameter* и *PyroParam* атрибутами *PyroSample*, которые задают *априорные распределения*:

```
linear.weight = PyroSample(dist.Normal(0, 1).expand([5, 2]).to_event(2))
```

- *Непараметрические вероятностные модели* – это класс статистических моделей, которые, в отличие от параметрических моделей, не предполагают фиксированной параметрической формы распределения данных.
- Такие модели обычно обладают *большей гибкостью*, поскольку они не ограничиваются заранее заданным набором параметров.
- **Случайный процесс** – это *семейство случайных величин* ($\{X_t, t \in T\}$), определенных на некотором вероятностном пространстве $(\Omega, \mathcal{F}, \mathrm{P})$.
- Случайный процесс можно представить как *случайную функцию*, которая для каждого фиксированного t выражает *случайную величину*, а для каждого *фиксированного* исхода $\omega \in \Omega$ *реализацию процесса* (*траекторию*), т.е. *функцию* $t \mapsto X(t, \omega)$.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

- Гауссов процесс является обобщением многомерного нормального распределения на *бесконечное* количество переменных.
- Он определяется как *распределение над функциями* $f(\mathbf{x})$, в котором для любого конечного набора точек $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ значения $f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_n)$ имеют совместное *многомерное нормальное распределение*:

$$f \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big) \quad \Leftrightarrow \quad p\big(f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_n)\big) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

- Гауссов процесс полностью определяется двумя характеристиками:
  - ✓ *Функцией среднего* $m(\mathbf{x})$, которая задает ожидаемое значение функции в точке $\mathbf{x}$:

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})].$$

  - ✓ *Ковариационной функцией* (или *ядром*) $k(\mathbf{x}, \mathbf{x}')$, которая задает ковариацию между значениями функции в точках $\mathbf{x}$ и $\mathbf{x}'$:

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\big[\big(f(\mathbf{x}) - m(\mathbf{x})\big)\big(f(\mathbf{x}') - m(\mathbf{x}')\big)\big].$$

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Постановка задачи регрессии для гауссовых процессов

- Постановка **задачи регрессии**:

  Дан *обучающий датасет*: $\mathcal{D} = (\mathbf{X}, \mathbf{y})$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$, $\mathbf{x}_i \in \mathbb{R}^d$;
  $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n]$, $\mathbf{y}_i \in \mathbb{R}$.

  Требуется для данных из *тестового датасета*: $\mathcal{D}_* = \mathbf{X}_* = [\mathbf{x}_{*1}, \mathbf{x}_{*2}, \dots, \mathbf{x}_{*m}]$:

  - *предсказать* значение $\mathbf{y}_* = f(\mathbf{X}_*)$;
  - *оценить неопределенность* этого предсказания.

- Будем решать задачу, полагая, что связь между признаками и целевой переменной задаётся с помощью *гауссова процесса*:

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big).$$

- Поскольку признаки всегда можно *центрировать*, то можно считать, что $m(\mathbf{x}) = 0$. Тогда гауссов процесс определяется *только ядром* $k(\mathbf{x}, \mathbf{x}')$.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Выбор ядра для гауссовых процессов

- Ядро $k(\mathbf{x}, \mathbf{x}')$ определяет *свойства моделируемой функции* $f(\mathbf{x})$, такие как гладкость, масштаб изменений и т.д.
- Одним из самых популярных ядер является *радиально-базисное ядро* (*RBF*, или *квадратичное экспоненциальное ядро*):

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

  где σ_f^2 – дисперсия (определяет амплитуду функции), l – параметр масштаба (определяет скорость изменения функции).
- *Линейное ядро* $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}'$ применяется для *обычной линейной регрессии*. Чтобы иметь возможность моделировать нелинейные зависимости, в гауссовских процессах используются *нелинейные ядра*, такие как RBF.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Совместное распределение обучающих и тестовых данных

- Согласно определению гауссова процесса, *совместное распределение* значений функции в *обучающих* и *тестовых* точках имеет вид:

$$\begin{bmatrix} \mathbf{y}(\mathbf{X}) \\ \mathbf{y}_*(\mathbf{X}_*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}(\mathbf{X},\mathbf{X}) + \sigma^2\mathbb{I} & \mathbf{K}(\mathbf{X},\mathbf{X}_*) \\ \mathbf{K}(\mathbf{X}_*,\mathbf{X}) & \mathbf{K}(\mathbf{X}_*,\mathbf{X}_*) \end{bmatrix} \right),$$

  - ✓ $\mathbf{K}(\mathbf{X},\mathbf{X}) \in \mathbb{R}^{n\times n}$ – *матрица ковариаций* между всеми парами обучающих точек с элементами: $[\mathbf{K}(\mathbf{X},\mathbf{X})]_{ij} = k(\mathbf{x}_i,\mathbf{x}_j)$;
  - ✓ $\mathbf{K}(\mathbf{X},\mathbf{X}_*) \in \mathbb{R}^{n\times m}$ – *матрица ковариаций* между обучающими и тестовыми точками с элементами: $[\mathbf{K}(\mathbf{X},\mathbf{X}_*)]_{ij} = k\left(\mathbf{x}_i,\mathbf{x}_{*j}\right)$;
  - ✓ $\mathbf{K}(\mathbf{X}_*,\mathbf{X}) = \mathbf{K}(\mathbf{X},\mathbf{X}_*)$, $\mathbf{K}(\mathbf{X}_*,\mathbf{X}) \in \mathbb{R}^{m\times n}$;
  - ✓ $\mathbf{K}(\mathbf{X}_*,\mathbf{X}_*) \in \mathbb{R}^{m\times m}$ – *матрица ковариаций* между всеми парами тестовых точек с элементами: $[\mathbf{K}(\mathbf{X}_*,\mathbf{X}_*)]_{ij} = k\left(\mathbf{x}_{*i},\mathbf{x}_{*j}\right)$;
  - ✓ $\sigma^2\mathbb{I}$ – *матрица шума*, добавляемая к ковариационной матрице.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Условное распределение для тестовых данных

- Для предсказания значений $\mathbf{y}_*$ в *тестовых точках* нужно найти *условное распределение*: $p(\mathbf{y}_* | \mathbf{y}, \mathbf{X}, \mathbf{X}_*)$.
- Из свойств *многомерного нормального распределения* известно, что условное распределение также является *нормальным*.
- *Можно показать*, что для полученного совместного распределения $p(\mathbf{y}_*, \mathbf{y})$ условное распределение $p(\mathbf{y}_* | \mathbf{y}, \mathbf{X}, \mathbf{X}_*)$ выражается следующим образом:

$$p(\mathbf{y}_* | \mathbf{y}, \mathbf{X}, \mathbf{X}_*) = \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*),$$

а *среднее значение* и *ковариационная матрица* задаются выражениями:

$$\boldsymbol{\mu}_* = \mathbf{K}(\mathbf{X}_*, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbb{I}]^{-1}\mathbf{y};$$

$$\boldsymbol{\Sigma}_* = \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbb{I}]^{-1}\mathbf{K}(\mathbf{X}, \mathbf{X}_*);$$

- Это выражение и задаёт *распределение предсказанных значений* $\mathbf{y}_*$.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Интерпретация результатов решения

- *Полученные значения* $\boldsymbol{\mu}_*$ и $\boldsymbol{\Sigma}_*$ можно *интерпретировать* таким образом:
  - ✓ *Среднее значение* $\boldsymbol{\mu}_*$ является наилучшим предсказанием $\mathbf{y}_*$. Если требуется точечное предсказание, то достаточно получить $\boldsymbol{\mu}_*$.
  - ✓ Ковариационная матрица содержит информацию о:
    - ➢ *Дисперсии* – диагональные элементы матрицы ковариации соответствуют дисперсии предсказания в каждой тестовой точке. Среднеквадратичное отклонение предсказания в точке $\mathbf{x}_{*i}$:

      $$\sigma(\mathbf{x}_{*i}) = \sqrt{[\boldsymbol{\Sigma}_*]_{ii}},$$

      Это позволяет строить доверительные интервалы вокруг $\boldsymbol{\mu}_*$.
    - ➢ *Ковариации* – недиагональные элементы $\boldsymbol{\Sigma}_*$ показывают, насколько предсказания в разных тестовых точках *коррелируют* друг с другом.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Определение гиперпараметров функции

- Приведённая функция содержит *неизвестные*: гиперпараметры ядра (σ_f^2 и l) и дисперсию шума (σ^2). Их требуется определить *из данных*.
- Для этого используется *метод максимального правдоподобия*.
- Так как $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbb{I})$, то *логарифм правдоподобия*:

$$\mathcal{L} = \log p\left(\mathbf{y} \mid \mathbf{X}, \sigma_f^2, l, \sigma^2\right) = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbb{I}]^{-1}\mathbf{y} - \frac{1}{2}\log\det(\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbb{I}) - \frac{n}{2}\log(2\pi)\,.$$

- Максимизация логарифма правдоподобия эквивалентна *минимизации отрицательного логарифма правдоподобия* по гиперпараметрам:

$$\sigma_f^{2^*}, l^*, \sigma^{2^*} = \underset{\left(\sigma_f^2, l, \sigma^2\right) \in \Theta}{\arg\min} \left[-\log p\left(\mathbf{y} \mid \mathbf{X}, \sigma_f^2, l, \sigma^2\right)\right].$$

- Решение данной задачи оптимизации позволит получить значения *гиперпараметров*, которые *используются в выражениях* для $\boldsymbol{\mu}_*$ и $\boldsymbol{\Sigma}_*$.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Ограничения гауссовских стохастических процессов

- Хотя гауссовы процессы являются весьма *мощными моделями*, у них есть несколько важных *ограничений*:

1. *Вычислительная сложность*. Обучение модели требует *инверсии* матрицы $\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbb{I}$, алгоритм которой имеет *сложность* $\mathcal{O}(n^3)$. Получение предсказаний для m тестовых точек имеет *сложность* $\mathcal{O}(nm + m^2)$, что ограничивает использование гауссовых процессов для *больших наборов данных*.
2. *Чувствительность к выбору ядра*. Результаты сильно зависят от выбора *ядра* и его *гиперпараметров*. Неправильный выбор может привести к *плохой аппроксимации данных*.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Ограничения гауссовских стохастических процессов

3. *Предположение о нормальности*. Гауссовы процессы предполагают, что данные подчиняются *нормальному распределению*. Если шум в данных не является нормальным (например, имеет тяжелые хвосты), модель может давать некорректные результаты.
4. *Сложности с экстраполяцией*. Гауссовы процессы *плохо работают* при *экстраполяции* за пределы области обучающих данных, особенно с RBF-ядром, так как предсказания стремятся к среднему значению (обычно нулю) с *увеличением расстояния от обучающих точек*.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Оптимизация гиперпараметров параметрических моделей

- Пусть дана параметрическая модель (например, *нейронная сеть*, *регрессия* или *градиентный бустинг*) с *гиперпараметрами* $\boldsymbol{\lambda} \in \Lambda$.
- Здесь Λ – *пространство гиперпараметров* (например, *скорость обучения*, *количество слоев*, *регуляризационные коэффициенты* и т.д.).
- Требуется найти такие *гиперпараметры* $\boldsymbol{\lambda}^*$, которые *минимизируют* некоторую целевую функцию $f(\boldsymbol{\lambda})$, например, *ошибку на валидационной выборке*:

$$\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda} \in \Lambda} f(\boldsymbol{\lambda}).$$

- При этом предполагается, что *поиск параметров* $\mathbf{w}$ параметрической модели $f(\boldsymbol{\lambda}) = f_{\mathbf{w}}(\boldsymbol{\lambda})$ осуществляется *в процессе обучения* модели для *фиксированных значений* $\boldsymbol{\lambda}$.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Проблемы определения гиперпараметров

- При определении гиперпараметров следует иметь в виду:
  1. Функция $f(\boldsymbol{\lambda})$ обычно *не имеет аналитического вида* (вычисление её значений возможно только после обучения модели $f_{\mathbf{w}}(\boldsymbol{\lambda})$ и оценки её производительности).
  2. Расчёт $f(\boldsymbol{\lambda})$ может быть *вычислительно сложным* (например, обучение нейронной сети).
  3. Пространство гиперпараметров Λ может быть сложным (*непрерывным*, *дискретным* или *смешанным*), а сама функция $f(\boldsymbol{\lambda})$ – *мультимодальной, шумной и недифференцируемой*.
- Для решения такой задачи используется **байесовская оптимизация**, *эффективно выполняющая глобальную оптимизацию «черного ящика»*.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Моделирование целевой функции с помощью гауссовых процессов

- Пусть целевая функция $f(\boldsymbol{\lambda}) \sim \mathcal{GP}\big(m(\boldsymbol{\lambda}), k(\boldsymbol{\lambda}, \boldsymbol{\lambda}')\big)$.
- Здесь $m(\boldsymbol{\lambda})$ – функция *среднего значения* (обычно принимается равной *нулю*), а $k(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ – *ковариационная функция* (например, *ядро RBF*).
- На каждом шаге t байесовской оптимизации имеется набор уже вычисленных значений функции:

$$\mathcal{D}_t = \{(\boldsymbol{\lambda}_i, f(\boldsymbol{\lambda}_i))\}_{i=1}^{t}$$

- Используя GP, можно вычислить *апостериорное распределение* $f(\boldsymbol{\lambda})$ в любой точке $\boldsymbol{\lambda}$:

$$p(f(\boldsymbol{\lambda}) \mid \mathcal{D}_t) = \mathcal{N}\Big(\mu_t(\boldsymbol{\lambda}), \sigma_t^2(\boldsymbol{\lambda})\Big),$$

где *среднее* и *дисперсия* вычисляются аналогично *формулам регрессии*.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Моделирование целевой функции с помощью гауссовых процессов

$$\mu_t(\boldsymbol{\lambda}) = \mathbf{k}_t(\boldsymbol{\lambda})^{\mathrm{T}}[\mathbf{K}_t + \sigma_n^2 \mathbb{I}]^{-1}\mathbf{f}_t;$$
$$\sigma_t^2(\boldsymbol{\lambda}) = k(\boldsymbol{\lambda}, \boldsymbol{\lambda}) - \mathbf{k}_t(\boldsymbol{\lambda})^{\mathrm{T}}[\mathbf{K}_t + \sigma_n^2 \mathbb{I}]^{-1}\mathbf{k}_t(\boldsymbol{\lambda}).$$

- Здесь:

  $\mathbf{f}_t = [f(\boldsymbol{\lambda}_1), f(\boldsymbol{\lambda}_2), \dots, f(\boldsymbol{\lambda}_t)]^{\mathrm{T}}$ – вектор *значений* целевой функции;

  $[\mathbf{K}_t]_{ij} = k(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}_j)$ – *ковариационная матрица* размером $t \times t$;

  $\mathbf{k}_t(\boldsymbol{\lambda}) = [k(\boldsymbol{\lambda}, \boldsymbol{\lambda}_1), k(\boldsymbol{\lambda}, \boldsymbol{\lambda}_2), \dots, k(\boldsymbol{\lambda}, \boldsymbol{\lambda}_t)]^{\mathrm{T}}$ – *вектор ковариаций* между точкой $\boldsymbol{\lambda}$ и точками из $\mathcal{D}_t$;

  $k(\boldsymbol{\lambda}, \boldsymbol{\lambda})$ – *априорная дисперсия* в точке $\boldsymbol{\lambda}$;

  σ_n^2 – *дисперсия шума*.

- Эти формулы позволяют не только *предсказать значение* $f(\boldsymbol{\lambda})$ в новой точке $\boldsymbol{\lambda}$, но и *оценить неопределенность предсказания* через $\sigma_t^2(\boldsymbol{\lambda})$.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Функция принятия решений (acquisition function)

- На каждой итерации байесовской оптимизации требуется выбрать *следующую точку* $\boldsymbol{\lambda}_{t+1}$, в которой будет вычислена *целевая функция* $f(\boldsymbol{\lambda})$.
- Но вычисление целевой функции *вычислительно сложно*, поэтому *перебор* всех возможных точек *не является эффективной* стратегией.
- Введём *функцию принятия решений* $a(\boldsymbol{\lambda})$, которая будет оценивать некоторую «*полезность*» вычисления $f(\boldsymbol{\lambda})$ в данной точке.
- Тогда *оптимальная* следующая точка определяется выражением:

$$\boldsymbol{\lambda}_{t+1} = \arg\max_{\boldsymbol{\lambda}\in\Lambda} a(\boldsymbol{\lambda}).$$

- Функция $a(\boldsymbol{\lambda})$ *зависит* от *апостериорного распределения* $p(f(\boldsymbol{\lambda}) \mid \mathcal{D}_t)$, т.е., от величин $\mu_t(\boldsymbol{\lambda})$ и $\sigma_t^2(\boldsymbol{\lambda})$. Существует *несколько популярных вариантов* функций принятия решений.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Функция принятия решений: ожидаемое улучшение

- Одна из функций принятия решений – *ожидаемое улучшение* (*Expected Improvement*, *EI*).
- Пусть f_{best} – *текущее лучшее значения* целевой функции:

$$f_{best} = \min_{i=1,2,\dots,t} f(\boldsymbol{\lambda}_i)\,.$$

- *Улучшение* в точке $\boldsymbol{\lambda}$ определяется выражением:

$$I(\boldsymbol{\lambda}) = \max(f_{best} - f(\boldsymbol{\lambda}), 0)\,.$$

- Так как $f(\boldsymbol{\lambda})$ – случайная величина, то вычисляем *ожидаемое улучшение*:

$$a_{EI}(\boldsymbol{\lambda}) = \mathbb{E}[I(\boldsymbol{\lambda})] = \mathbb{E}[\max(f_{best} - f(\boldsymbol{\lambda}), 0)].$$

- Для *нормального распределения* ожидаемое улучшение может быть выражено *аналитически*.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Функция принятия решений: вероятность улучшения

- Другая функция принятия решений – *вероятность улучшения (Probability of Improvement, PI)*.
- Она измеряет *вероятность* того, что $f(\boldsymbol{\lambda})$ будет лучше, чем f_{best}:

$$a_{PI}(\boldsymbol{\lambda}) = \mathrm{P}(f(\boldsymbol{\lambda}) < f_{best}) = \Phi\left(\frac{f_{best} - \mu_t(\boldsymbol{\lambda})}{\sigma_t(\boldsymbol{\lambda})}\right).$$

- Здесь $\Phi(\cdot)$ – функция распределения стандартного нормального распределения.
- Величина PI *проще в вычислении*, но *менее эффективна*, чем EI, так как она не учитывает *величину улучшения*, а только *вероятность его достижения*.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Оптимизация функции принятия решений

- После выбора *функции принятия решений* нужно найти:
$$\boldsymbol{\lambda}_{t+1} = \arg\max_{\boldsymbol{\lambda}\in\Lambda} a(\boldsymbol{\lambda}) .$$
- Данная задача оптимизации ищет максимум функции принятия решений, которая, в отличие от *целевой функции*, относительно *легко вычисляется*, поскольку зависит *только от величин* $\mu_t(\boldsymbol{\lambda})$ и $\sigma_t^2(\boldsymbol{\lambda})$. А они вычисляются *аналитически* через GP.
- Для *оптимизации* обычно используются *градиентный спуск*, *L-BFGS* и т.д.
- После получения $\boldsymbol{\lambda}_{t+1}$ нужно *вычислить* $f(\boldsymbol{\lambda}_{t+1})$ и добавить результат в $\mathcal{D}_{t+1}$, а затем *повторить* процесс до достижения *критерия остановки*.
- В результате получаются *оптимальные гиперпараметры* $\boldsymbol{\lambda}^*$:
$$\boldsymbol{\lambda}^* = \arg\min_{\boldsymbol{\lambda}_i\in\{\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2,\ldots,\boldsymbol{\lambda}_T\}} f(\boldsymbol{\lambda}_i) .$$

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Демонстрация практических примеров

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Заключение

1. Поговорили про недостатки линейных моделей, узнали про обобщенные линейные модели, ввели понятие функции связи.
2. Вспомнили экспоненциальное семейство распределений, выяснили, как составлять обобщенную линейную модель для различных распределений.
3. Разобрали примеры обобщенных линейных моделей.
4. Ввели понятие моделей смесей, определили структуру моделей смесей.
5. Рассмотрели часто используемые примеры моделей смесей.
6. Ввели понятия байесовской и стохастической нейронных сетей.
7. Рассмотрели непараметрические модели, поговорили о гауссовых процессах и основных задачах, решаемых с их помощью.
8. На практических примерах посмотрели реализацию моделей в Pyro.

VOLGOGRAD STATE TECHNICAL UNIVERSITY

Спасибо за внимание!

Волгоград 2025