Нейронные сети. Краткий курс.

Лекция 5

Обучение сети RBF как плохо обусловленная задача.

Описанная в предыдущей лекции процедура обучения нейронной сети на основе радиальных базисных функций (далее RBF-сеть) не всегда работает. Множество обучающих точек может быть очень большим, матрица интерполяции $\boldsymbol{\Phi}$ может оказаться плохо обусловленной, множество примеров может неудачно покрывать пространство и все это может привести к плохой обобщающей способности сети.

Здесь мы сталкиваемся с типом задач, которые называются *некорректными или плохо обусловленными.* Чтобы прочувствовать, что значит плохо обусловленная задача, рассмотрим для примера простую систему линейных уравнений

$$\begin{aligned} x_1 + 10\ x_2 &= 11 \\ 10\ x_1 + 101\ x_2 &= 111 \end{aligned}$$

Или то же в матричной форме

$$\begin{pmatrix} 1 & 10 \\ 10 & 101 \end{pmatrix} \times \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 11 \\ 111 \end{pmatrix}$$

Решением этой системы является точка на плоскости с координатами $x_1 = 1,\ x_2 = 1$. Теперь представим, что в результате измерительной неточности при определении правой части обусловленной шумами система слегка видоизменилась:

$$\begin{pmatrix} 1 & 10 \\ 10 & 101 \end{pmatrix} \times \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 11.1 \\ 111 \end{pmatrix} \quad .$$

И решение этой «испорченной» шумами системы уже $x_1 = 11.1,\ x_2 = 0$! Довольно далеко от решения «точной» системы. В матричной форме эта простая задачка представлена чтобы вызвать ассоциации с системой (4.12) из предыдущей лекции.

Итак рассмотрим неизвестное отображение $f : \boldsymbol{x} \in \boldsymbol{X} \to y \in \boldsymbol{Y},\ \boldsymbol{X} \subset \mathbb{R}^{m_0}, \boldsymbol{Y} \subset \mathbb{R}$, то есть отображающее область $\boldsymbol{X}$ в диапазон $\boldsymbol{Y}$, что собственно и призвана делать рассматриваемая нами RBF-сеть. Задача реконструкции отображения f считается хорошо обусловленной (или корректной) если выполнены следующие три условия.

1. *Существование*. Для любого входного вектора $\boldsymbol{x} \in \boldsymbol{X}$ существует выходное значение $y = f(\boldsymbol{x})$, где $y \in \boldsymbol{Y}$.
2. *Однозначность*. Для любой пары векторов $\boldsymbol{x}, \boldsymbol{t} \in \boldsymbol{X}$ равенство $f(\boldsymbol{x}) = f(\boldsymbol{t})$ выполняется тогда и только тогда, когда $\boldsymbol{x} = \boldsymbol{t}$.
3. *Непрерывность*. Отображение считается непрерывным, если для любого $\epsilon > 0$ существует $\delta(\epsilon) > 0$, такое, что из условия $\rho_x(\boldsymbol{x}, \boldsymbol{t}) < \epsilon$ вытекает условие $\rho_y(f(\boldsymbol{x}), f(\boldsymbol{t})) < \delta$, где ρ - расстояние в соответствующих пространствах. Свойство непрерывности еще называют *устойчивостью*.

Если какое-то из этих условий не выполнено, задача считается плохо обусловленной. В обучающем наборе данных информации может оказаться недостаточно для реконструкции отображения, шумы могут оказаться достаточно большими, чтобы вывести результат отображения из диапазона $\boldsymbol{Y}$, что будет означать нарушение первого условия. Собственно сама задача может оказаться таковой, что является неустойчивой относительно возмущений в выходных данных. Рассмотрим далее как с помощью методов *регуляризации* плохо обусловленную задачу сделать хорошо обусловленной. Обычно теорию регуляризации связывают с именем Тихонова А.Н., который в 1963 г. предложил новый метод названный регуляризацией и предназначенный для решения плохо обусловленных задач. Главная идея регуляризации заключается в стабилизации решения с помощью некоторой вспомогательной неотрицательной функции, которая несет в себе априорную информацию о решении. Наиболее общей формой априорной информации является предположение о гладкости искомого отображения.

Пусть для создания RBF-сети (или реконструкции отображения) имеется множество пар данных $(\boldsymbol{x}_i, d_i)$ доступных для аппроксимации, где $\boldsymbol{x}_i \in \mathbb{R}^{m_0}$, $i = 1, 2, \ldots, N$ и $y_i \in \mathbb{R}$, $i = 1, 2, \ldots, N$. Функцию аппроксимации обозначим как $F(\boldsymbol{x})$, где для простоты обозначений опущен вектор весов $\boldsymbol{w}$. В теории регуляризации Тихонова функция ошибок, которую нужно минимизировать для определения параметров искомого отображения модифицируется следующим образом:

$$E(F) = \frac{1}{2}\sum_{i=1}^{N}(d_i - y_i)^2 + \frac{1}{2}\lambda\|\boldsymbol{D}F\| = E_s(F) + \lambda E_c(F) \quad . \tag{5.1}$$

Здесь первое слагаемое это *слагаемое стандартной ошибки* описывающее расстояние между желаемым откликом и фактическим выходным сигналом, а второе слагаемое это *слагаемое регуляризации* в котором $\boldsymbol{D}$ - линейный дифференциальный оператор, называемый иногда *стабилизатором* решения, содержащий априорную информацию о задаче. Параметр λ - действительное положительное число, называемое *параметром регуляризации*. Обозначим аргминимум функционала Тихонова $E(F)$ как $F_\lambda(\boldsymbol{x})$. Параметр регуляризации λ можно рассматривать как индикатор достаточности имеющегося набора данных для определения решения $F_\lambda(\boldsymbol{x})$. Крайний случай $\lambda \to 0$ означает, что задача является хорошо обусловленной и имеет решение целиком зависящее от набора учебных данных.

Другой крайний случай большого параметра регуляризации означает, что самого априорного ограничения на гладкость определяемого оператором $\boldsymbol{D}$ достаточно для определения решения или что учебные данные содержат недостаточно информации для определения отображения. В практических приложениях параметр регуляризации подбирается между этими крайними случаями. Можно рассматривать функцию $E_c(F)$ как *функцию штрафа за сложность модели*, влияние которой определяется параметром регуляризации.

Итак, принцип регуляризации заключается в следующем. Необходимо найти функцию $F(\boldsymbol{x})$ минимизирующую функционал Тихонова (5.1). Чтобы минимизировать функционал, нужно задать правило оценки его дифференциала. Для этого используем *дифференциал Фреше*:

$$dE(F,h) = \frac{d}{d\beta} E(F(\boldsymbol{x})+\beta h(\boldsymbol{x}))_{\beta=0} , \tag{5.2}$$

где $h(\boldsymbol{x})$ - фиксированная функция. Необходимым условием того, чтобы функция $F(\boldsymbol{x})$ была экстремумом функционала $E(F)$ является равенство

$$dE(F,h) = dE_s(F,h)+\lambda dE_c(F,h) = 0 , \tag{5.3}$$

где dE_s и dE_c - соответствующие дифференциалы Фреше слагаемых стандартной ошибки и регуляризации. Дифференцирование в определении дифференциала Фреше осуществляется по обычным правилам. Рассмотрим первое слагаемое в (5.3):

$$\begin{aligned} dE_s(F,h) &= \frac{1}{2}\frac{d}{d\beta}\sum_{i=1}^{N}(d_i-F(\boldsymbol{x}_i)-\beta h(\boldsymbol{x}_i))^2_{\beta=0} = \\ &= -\sum_{i=1}^{N}(d_i-F(\boldsymbol{x}_i)-\beta h(\boldsymbol{x}_i))h(\boldsymbol{x}_i)_{\beta=0} = -\sum_{i=1}^{N}(d_i-F(\boldsymbol{x}_i))h(\boldsymbol{x}_i) \end{aligned} . \tag{5.4}$$

Нам понадобится следующая теорема Ритца о представлении из функционального анализа. *Пусть* f *- линейный ограниченный функционал в Гильбертовом пространстве* $\boldsymbol{H}$. *Тогда существует единственная функция* $h_0 \in \boldsymbol{H}$ *такая что* $f=(h,h_0)$ *для всех* $h \in \boldsymbol{H}$ *, при этом* $\|f\|_{\tilde{\boldsymbol{H}}}=\|h_0\|_{\boldsymbol{H}}$ *, где* $\tilde{\boldsymbol{H}}$ *- пространство сопряженное пространству* $\boldsymbol{H}$. В связи с этим можно переписать (5.4) в виде скалярного произведения в функциональном пространстве:

$$dE_s(F,h) = -\left(h,\ \sum_{i=1}^{N}(d_i-F(\boldsymbol{x}_i))\delta(\boldsymbol{x}-\boldsymbol{x}_i)\right) , \tag{5.5}$$

где $\delta(\boldsymbol{x}-\boldsymbol{x}_i)$ - дельта функция Дирака. Пользуясь подобным представлением, запишем выражение для дифференциала Фреше для dE_c :

$$dE_c(F,h) = \frac{d}{d\beta} E_c(F(\boldsymbol{x})+\beta h(\boldsymbol{x}))_{\beta=0} = \frac{1}{2}\frac{d}{d\beta}\int_{\mathbb{R}_{m_0}} (\boldsymbol{D}[F(\boldsymbol{x})+\beta h(\boldsymbol{x})])^2 d\boldsymbol{x}_{\beta=0} =$$
$$= \int_{\mathbb{R}_{m_0}} \boldsymbol{D}[F(\boldsymbol{x})+\beta h(\boldsymbol{x})]\boldsymbol{D}h(\boldsymbol{x})d\boldsymbol{x}_{\beta=0} = \int_{\mathbb{R}_{m_0}} \boldsymbol{D}F(\boldsymbol{x})\boldsymbol{D}h(\boldsymbol{x})d\boldsymbol{x} = \quad (5.6)$$
$$= \left(\boldsymbol{D}F(\boldsymbol{x}),\ \boldsymbol{D}h(\boldsymbol{x})\right).$$

Для данного оператора можно найти такой однозначно определенный оператор $\tilde{\boldsymbol{D}}$ называемый *сопряженным*, такой что для любых функций $u(\boldsymbol{x})$ и $v(\boldsymbol{x})$ удовлетворяющих соответствующим граничным условиям, можно записать

$$\int_{\mathbb{R}^{m_0}} u(\boldsymbol{x})\boldsymbol{D}v(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathbb{R}^{m_0}} v(\boldsymbol{x})\tilde{\boldsymbol{D}}u(\boldsymbol{x})d\boldsymbol{x} \quad . \quad (5.7)$$

Это тождество называют тождеством Грина. Если теперь переобозначить $u(\boldsymbol{x})=\boldsymbol{D}F(\boldsymbol{x})$ и $\tilde{\boldsymbol{D}}v(\boldsymbol{x})=\boldsymbol{D}h(\boldsymbol{x})$, то можно записать следующее выражение:

$$dE_c(F,h) = \int_{\mathbb{R}^{m_0}} \boldsymbol{D}F(\boldsymbol{x})\boldsymbol{D}h(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathbb{R}^{m_0}} h(\boldsymbol{x})\tilde{\boldsymbol{D}}\boldsymbol{D}F(\boldsymbol{x})d\boldsymbol{x} = \left(h,\ \tilde{\boldsymbol{D}}\boldsymbol{D}F(\boldsymbol{x})\right) \quad (5.8)$$

Если теперь переписать условие экстремума функционала Тихонова в виде $\frac{1}{\lambda}dE_s+dE_c=0$ и подставить туда выражения для дифференциалов Фреше, то получим

$$-\frac{1}{\lambda}\left(h(\boldsymbol{x}),\ \sum_{i=1}^{N}(d_i-F(\boldsymbol{x}))\delta(\boldsymbol{x}-\boldsymbol{x}_i)\right) + \left(h(\boldsymbol{x}),\ \tilde{\boldsymbol{D}}\boldsymbol{D}F(\boldsymbol{x})\right) = 0 \quad . \quad (5.9)$$

Условие (5.9) можно переписать в виде одного скалярного произведения:

$$\left(h(\boldsymbol{x}),\ \tilde{\boldsymbol{D}}\boldsymbol{D}F(\boldsymbol{x})-\frac{1}{\lambda}\sum_{i=1}^{N}(d_i-F(\boldsymbol{x}_i))\delta(\boldsymbol{x}-\boldsymbol{x}_i)\right) = 0 \quad . \quad (5.10)$$

Условие (5.10) будет выполняться для произвольной функции $h(\boldsymbol{x})$ если

$$\tilde{\boldsymbol{D}}\boldsymbol{D}F(\boldsymbol{x}) = \frac{1}{\lambda}\sum_{i=1}^{N}(d_i-F(\boldsymbol{x}))\delta(\boldsymbol{x}-\boldsymbol{x}_i) \quad - \quad (5.11)$$

уравнение Эйлера-Лагранжа для функционала Тихонова $E(F)$. Оно является необходимым условием существования экстремума для функции $F_\lambda(\boldsymbol{x})$. Уравнение (5.11) это дифференциальное уравнение в частных производных для функции аппроксимации F .

Для заданного линейного дифференциального оператора $\boldsymbol{L}$, зададимся функцией $G(\boldsymbol{x},\boldsymbol{\xi})$ обладающей следующими свойствами:

1). Для фиксированного значения $\boldsymbol{\xi}$ функция $G(\boldsymbol{x},\boldsymbol{\xi})$ удовлетворяет граничным условиям.

2). Во всех точках, кроме $\boldsymbol{x}=\boldsymbol{\xi}$, все производные функции $G(\boldsymbol{x},\boldsymbol{\xi})$ при фиксированном $\boldsymbol{\xi}$ непрерывны, а их количество определяется порядком $\boldsymbol{L}$.

3). $G(\boldsymbol{x},\boldsymbol{\xi})$ как функция $\boldsymbol{x}$ удовлетворяет уравнению в частных производных

$$\boldsymbol{L}\,G(\boldsymbol{x},\boldsymbol{\xi})=\delta(\boldsymbol{x}-\boldsymbol{\xi}) \quad . \tag{5.12}$$

Такая функция называется *функцией Грина* для оператора дифференцирования $\boldsymbol{L}$. Она играет ту же роль для $\boldsymbol{L}$, что и обратная матрица в матричном исчислении.

Пусть $\phi(\boldsymbol{x})$ - непрерывная или кусочно-непрерывная функция. Тогда функция $F(\boldsymbol{x})$, определенная как

$$F(\boldsymbol{x}) \;=\; \int\limits_{\mathbb{R}^{m_0}} G(\boldsymbol{x},\boldsymbol{\xi})\,\phi(\boldsymbol{\xi})\,d\boldsymbol{\xi} \tag{5.13}$$

является решением дифференциального уравнения

$$\boldsymbol{L}\,F(\boldsymbol{x}) \;=\; \phi(\boldsymbol{x}) \quad . \tag{5.14}$$

Проверка этого утверждения выполняется непосредственно

$$\boldsymbol{L}\,F(\boldsymbol{x})=\boldsymbol{L}\int\limits_{\mathbb{R}^{m_0}} G(\boldsymbol{x},\boldsymbol{\xi})\,\phi(\boldsymbol{x})\,d\boldsymbol{\xi}=\int\limits_{\mathbb{R}^{m_0}} \boldsymbol{L}\,G(\boldsymbol{x},\boldsymbol{\xi})\,\phi(\boldsymbol{x})\,d\boldsymbol{\xi}=\int\limits_{\mathbb{R}^{m_0}} \delta(\boldsymbol{x}-\boldsymbol{\xi})\,\phi(\boldsymbol{x})\,d\boldsymbol{\xi}=\phi(\boldsymbol{x}) \quad . \tag{5.15}$$

Перестановка интеграла и оператора дифференцирования возможно потому, что дифференцирование и интегрирование производятся по разным переменным, дифференцирование по $\boldsymbol{x}$, а интегрирование по $\boldsymbol{\xi}$.

Вернемся к уравнению Эйлера-Лагранжа. В нем в качестве оператора $\boldsymbol{L}$ выступает оператор $\tilde{\boldsymbol{D}}\boldsymbol{D}$, а в качестве функции ϕ функция

$$\phi(\boldsymbol{\xi}) \;=\; \frac{1}{\lambda}\sum_{i=1}^{N}(d_i-F(\boldsymbol{x}_i))\,\delta(\boldsymbol{\xi}-\boldsymbol{x}_i) \quad . \tag{5.16}$$

Тогда запись $F(\boldsymbol{x})$ через функцию Грина выглядит следующим образом:

$$\begin{aligned} F_{\lambda}(\boldsymbol{x}) \;&=\; \int\limits_{\mathbb{R}^{m_0}} G(\boldsymbol{x},\boldsymbol{\xi})\left\{\frac{1}{\lambda}\sum_{i=1}^{N}(d_i-F(\boldsymbol{x}_i))\,\delta(\boldsymbol{\xi}-\boldsymbol{x}_i)\right\} \;= \\ &=\; \frac{1}{\lambda}\sum_{i=1}^{N}(d_i-F(\boldsymbol{x}_i))\int\limits_{\mathbb{R}^{m_0}} G(\boldsymbol{x},\boldsymbol{\xi})\,\delta(\boldsymbol{\xi}-\boldsymbol{x}_i)\,d\boldsymbol{\xi} \;= \\ &=\; \frac{1}{\lambda}\sum_{i=1}^{N}(d_i-F(\boldsymbol{x}))\,G(\boldsymbol{x},\boldsymbol{x}_i) \end{aligned} \tag{5.17}$$

А это означает, что решение задачи регуляризации является линейной суперпозицией N функций Грина с весами $\frac{1}{\lambda}(d_i - F(\boldsymbol{x}_i))$. Обозначив $w_i = \frac{1}{\lambda}(d_i - F(\boldsymbol{x}_i))$ можно записать (5.17) как

$$F_\lambda(\boldsymbol{x}) = \sum_{i=1}^{N} w_i G(\boldsymbol{x}, \boldsymbol{x}_i) \quad . \tag{5.18}$$

Вычисляя $F_\lambda(\boldsymbol{x})$ в точках $\boldsymbol{x}_j$, $j = 1,2,...,N$ можно получить систему уравнений

$$F_\lambda(\boldsymbol{x}_j) = \sum_{i=1}^{N} w_i G(\boldsymbol{x}_j, \boldsymbol{x}_i) \qquad j = 1,2,...,N \tag{5.19}$$

или то же самое в матричной форме

$$\boldsymbol{F}_\lambda = \boldsymbol{G}\boldsymbol{w} \quad , \tag{5.20}$$

где $\boldsymbol{F}_\lambda = (F_\lambda(\boldsymbol{x}_1), F_\lambda(\boldsymbol{x}_2), ..., F_\lambda(\boldsymbol{x}_N))^T$, $\boldsymbol{w} = \frac{1}{\lambda}(\boldsymbol{d} - \boldsymbol{F}_\lambda)$, $\boldsymbol{d} = (d_1, d_2, ..., d_N)^T$,

$$\boldsymbol{G} = \begin{pmatrix} G(\boldsymbol{x}_1, \boldsymbol{x}_1) & ... & G(\boldsymbol{x}_1, \boldsymbol{x}_N) \\ ... & ... & ... \\ G(\boldsymbol{x}_N, \boldsymbol{x}_1) & ... & G(\boldsymbol{x}_N, \boldsymbol{x}_N) \end{pmatrix} .$$

Используя выражение для вектора весов $\boldsymbol{w}$, переписав его как $\boldsymbol{F}_\lambda + \lambda \boldsymbol{w} = \boldsymbol{d}$, затем подставив вместо $\boldsymbol{F}_\lambda$ правую часть (5.20), можно получить следующее уравнение для определения весов:

$$(\boldsymbol{G} + \lambda \boldsymbol{I})\boldsymbol{w} = \boldsymbol{d} \quad , \tag{5.21}$$

где $\boldsymbol{I}$ - единичная матрица. Матрица $\boldsymbol{G}$ называется *матрицей Грина*, она аналогична матрице интерполяции $\boldsymbol{\Phi}$ и так же является симметричной в силу того, что мы определили оператор $\boldsymbol{L} = \tilde{\boldsymbol{D}}\boldsymbol{D}$ как самосопряженный, а это соответствует симметрии функции Грина, то есть $G(\boldsymbol{x}_j, \boldsymbol{x}_i) = G(\boldsymbol{x}_i, \boldsymbol{x}_j)$ для всех $i, j = 1,2,...,N$. Симметрия $\boldsymbol{G}$ еще означает, что $\boldsymbol{G}^T = \boldsymbol{G}$. На практике всегда можно подобрать λ , так чтобы матрица $(\boldsymbol{G} + \lambda \boldsymbol{I})$ была достаточно хорошей для ее обращения, следовательно система (5.21) имеет единственное решение

$$\boldsymbol{w} = (\boldsymbol{G}+\lambda \boldsymbol{I})^{-1}\boldsymbol{d} \quad . \tag{5.22}$$

Итак, выбрав оператор дифференцирования $\boldsymbol{D}$ и имея для него набор функций Грина можно получить набор весов $\boldsymbol{w}$ для заданного λ . Решение задачи регуляризации задается разложением (5.18) в котором функция Грина соответствует самосопряженному оператору $\tilde{\boldsymbol{D}}\boldsymbol{D}$. Характеристики функции Грина зависят от выбранной формы оператора $\boldsymbol{D}$ и выбранных граничных условий. Количество функций Грина в таком решении равно числу примеров обучения. Чтобы функция Грина оказалась еще и радиальной функцией нужно, чтобы оператор $\boldsymbol{D}$ был инвариантен к преобразованием связанным с переносом координат (тогда $G(\boldsymbol{x},\boldsymbol{x}_i)=G(\boldsymbol{x}-\boldsymbol{x}_i)$) и инвариантен к поворотам. В этом случае $G(\boldsymbol{x},\boldsymbol{x}_i)=G(\|\boldsymbol{x}-\boldsymbol{x}_i\|)$. В этом случае решение

$$F_\lambda(\boldsymbol{x}) = \sum_{i=1}^{N} w_i G(\|\boldsymbol{x}-\boldsymbol{x}_i\|) \tag{5.23}$$

очень похоже на решение задачи интерполяции, однако оно регуляризовано и будет совпадать с интерполяционным только в случае $\lambda=0$.

Такая функция Грина, которая одновременно является радиальной представляет практический интерес для построения RBF-сетей. Примером такой функции является многомерная функция Гаусса:

$$G(\boldsymbol{x},\boldsymbol{x}_i) = \exp\left(-\frac{1}{2\sigma_i^2}\|\boldsymbol{x}-\boldsymbol{x}_i\|^2\right) \quad . \tag{5.24}$$

Такую функцию Грина определяет оператор

$$\boldsymbol{L} = \sum_{n=0}^{\infty}(-1)^n \alpha_n \nabla^{2n} \quad , \tag{5.25}$$

где $\alpha_n=\frac{\sigma_i^{2n}}{n!2^n}$, ∇^{2n} - n -кратное применение оператора Лапласа в $\mathbb{R}^{m_0}$

$$\nabla^2 = \frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial x_2^2}+\ldots+\frac{\partial^2}{\partial x_{m_0}^2} \quad . \tag{5.26}$$