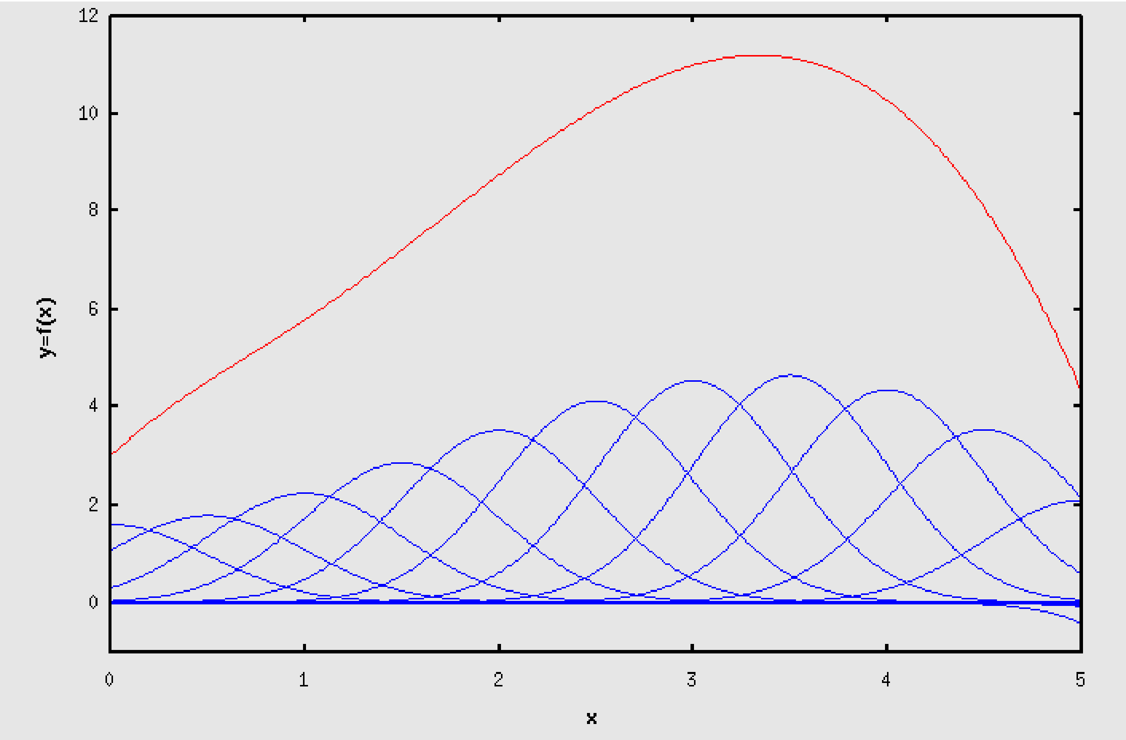
Нейронные сети. Краткий курс.

Лекция 4

Сети на основе радиальных базисных функций

Многослойный персептрон, рассмотренный в предыдущих лекциях выполняет аппроксимацию стохастической функции нескольких переменных путем преобразования множества входных переменных $\boldsymbol{x} \in \mathbb{R}^N$ во множество выходных переменных $\boldsymbol{y} \in \mathbb{R}^M$. В такой сети осуществляется аппроксимация *глобального* типа, это означает, что при формировании выходного сигнала участвуют выходные сигналы многих или даже всех нейронов. Другой способ отображения входного множества в выходное заключается в преобразовании основанном на нескольких одиночных аппроксимирующих функциях каждая из которых реализует ожидаемые значения только ограниченной области многомерного пространства. Самая простая аналогия это аппроксимация произвольной одномерной кривой суммой взвешенных гауссоид, см. рисунок ниже.



На этом рисунке красная кривая представлена суммой синих гауссоид, ясно что для некоторой точки x основной вклад дают лишь несколько гауссоид, центры которых близки к этой точке. Поэтому такая аппроксимация называется *локальной*. Можно сказать, что все преобразование в целом представляет собой сумму локальных преобразований на основе базисных функций (гауссоид в нашем примере).

В сетях с радиальными базисными функциями срытые нейроны реализуют функции радиально изменяющиеся вокруг выбранного центра и принимающие ненулевые значения только в окрестности этого центра. Подобные функции, определяемые в виде

$$\phi_i(\boldsymbol{x}) = \phi(\|\boldsymbol{x}-\boldsymbol{c}_i\|) \tag{4.1}$$

называются радиальными базисными функциями. В таких сетях роль скрытого нейрона заключается в отображении радиального пространства вокруг одиночной заданной точки либо вокруг группы точек образующих кластер. Суперпозиция сигналов от скрытых нейронов выполняемая выходным нейроном, позволяет получить отображения всей заданной области многомерного пространства.

Сигмоидальный нейрон можно представить в многомерном пространстве гиперплоскостью, которая разделяет это пространство на два класса в которых выполняется одно из двух условий: либо $\sum_{i=1}^{N} w_{ji} x_i > 0$ либо $\sum_{i=1}^{N} w_{ji} x_i < 0$. Радиальный же нейрон разделяет пространство гиперсферой вокруг центральной точки и осуществляет шаровое разделение пространства. В отличие от многослойного персептрона, сеть с радиальными базисными функциями имеет как правило только один скрытый слой с числом нейронов существенно превышающим число входов сети. Выходной нейрон осуществляет суммирование сигналов, генерируемых скрытыми нейронами.

Математические основы

Математическую основу функционирования сетей с радиальными базисными функциями (RBF-сетей) составляет теорема Ковера о разделимости образов, которая утверждает что *нелинейное преобразование сложной задачи классификации образов в пространство более высокой размерности повышает вероятность линейной разделимости образов*. Рассмотрим для начала что такое *разделимость образов*.

Рассмотрим семейство поверхностей, каждая из которых делит входное пространство на 2 части. Пусть $\boldsymbol{X}$ множество, состоящее из *N* образов (векторов) $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, каждый из которых принадлежит одному из 2-х классов. Такая *дихотомия* или бинарное разделение точек называется разделимой по отношению к семейству поверхностей, если в этом семействе существует поверхность, которая отделяет точки класса $\boldsymbol{X}_1$ от точек класса $\boldsymbol{X}_2$. Для каждого образа $\boldsymbol{x} \in \boldsymbol{X}$ определим вектор, состоящий из множества действительнозначных функций вида

$$\boldsymbol{\phi}(\boldsymbol{x}) = \left[\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \ldots, \phi_{m_1}(\boldsymbol{x})\right]^T . \tag{4.2}$$

Допустим m_0 это размерность векторов $\boldsymbol{x}$, тогда векторная функция $\boldsymbol{\phi}(\boldsymbol{x})$

отображает точки m_0 -мерного входного пространства в пространство размерности m_1 . Функции $\phi_i(\boldsymbol{x})$ называют *скрытыми*, поскольку они играют роль функций активации скрытых нейронов в сетях прямого распространения. Пространство образованное множеством скрытых функций называют *скрытым пространством* или *пространством признаков*.

Дихотомия $\{\boldsymbol{X}_1, \boldsymbol{X}_2\}$ множества $\boldsymbol{X}$ называется ϕ -разделимой (или ϕ -сепарабельной), если существует m_1 -мерный вектор $\boldsymbol{w}$, для которого можно записать

$$\begin{aligned} \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) > 0, \quad \boldsymbol{x} \in \boldsymbol{X}_1 \ , \\ \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) < 0, \quad \boldsymbol{x} \in \boldsymbol{X}_2 \ . \end{aligned} \tag{4.3}$$

Гиперплоскость

$$\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) \ = \ 0 \tag{4.4}$$

описывает поверхность в скрытом пространстве. Обратный образ этой поверхности

$$\boldsymbol{x}: \ \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) = 0 \tag{4.5}$$

определяет разделяющую поверхность во входном пространстве.

Рассмотрим *рациональные многообразия* *r*-го порядка во входном пространстве размерности m_0 , то есть гиперповерхности описываемые следующим уравнением порядка *r* в координатах входного вектора $\boldsymbol{x}$:

$$\sum_{0 \le i_1 \le i_2 \le \ldots \le i_r \le m_0} a_{i_1 i_2 \ldots i_r} x_{i_1} x_{i_2} \ldots x_{i_r} \ = \ 0 \ \ , \tag{4.6}$$

где вектор $\boldsymbol{x}$ дополнен как обычно компонентом $x_0 = 1$. Для входного пространства размерности m_0 сумма (4.6) содержит $\dfrac{(m_0 - r)!}{m_0! \, r!}$ слагаемых содержащих всевозможные сочетания из r множителей $x_{i_1} x_{i_2} \cdots x_{i_r}$ или *одночленов*.

В вероятностном эксперименте разделимость образов это случайное событие зависящее от выбранной дихотомии и вероятностного распределения образов во входном пространстве. Пусть образы $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ выбираются независимо, в соответствии с вероятностным распределением, присущим входному пространству, а все возможные дихотомии $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ равновероятны. Пусть $P(N, m_1)$ - вероятность того, что случайно выбранная дихотомия является ϕ -разделимой, если класс разделяющих гиперповерхностей имеет m_1 степеней свободы. Тогда

$$P(N, m_1) = \left(\frac{1}{2}\right)^{(N-1)} \sum_{m=0}^{m_1-1} \binom{N-1}{m} \quad , \tag{4.7}$$

где $\binom{N-1}{m}$ это биномиальный коэффициент, определяемый по формуле $\binom{l}{m} = \frac{l(l-1)...(l-m+1)}{m!}$ для всех целых l и m . Уравнение (4.7) отражает сущность *теоремы Ковера о разделимости* случайных образов. Собственно (4.7) это вероятность того, что $(N-1)$ подбрасываний монеты приведет к выпадению не более (m_1-1) решек. Ясно, что чем больше размерность скрытого пространства m_1 , тем ближе вероятность $P(N, m_1)$ к единице. Отметим два основных момента вытекающих из этой теоремы.

1. Необходимо определить скрытую функцию $\phi_i(\boldsymbol{x}), \quad i=1,2,...,m_1$.
2. Более высокая размерность скрытого пространства по сравнению со входным. Эта размерность определяется числом скрытых нейронов $m-1$.

В некоторых случаях однако, удается добиться линейной разделимости достаточно нелинейного преобразования $\boldsymbol{\phi}(\boldsymbol{x})$ без повышения размерности скрытого пространства.

Задача интерполяции

Итак, теперь рассмотрим сеть прямого распространения с одним входным, одним скрытым и одним выходным слоем, содержащим единственный нейрон для простоты выкладок, что однако не снижает общности рассмотрения. Такая сеть осуществляет *нелинейное отображение* входного пространства в скрытое, а затем *линейное* отображение скрытого пространства в выходное. В целом сеть реализует отображение

$$s \; : \; \mathbb{R}^{m_0} \rightarrow \mathbb{R}^1 \quad . \tag{4.8}$$

Отображение s можно рассматривать как гиперповерхность или график $\Gamma \subset \mathbb{R}^{(m_0+1)}$ изменения выходного сигнала в зависимости от входного. В общем случае эта поверхность неизвестна, а на данные обучения ещё накладывается шум. На *этапе обучения* нейронной сети поверхность Γ оптимизируется на основании известных точек данных представляемых сети в форме примеров содержащих как входные, так и им соответствующие выходные данные. Фаза *обобщения* равносильна интерполяции на интервалах между точками представленными в учебном наборе данных. Поверхность Γ после обучения точно проходит через все точки примеров обучения. Метод *радиальных базисных функций* сводится к подбору функции F , имеющей следующий вид:

$$F(\boldsymbol{x}) = \sum_{i=1}^{N} w_i \phi(\|\boldsymbol{x}-\boldsymbol{x}_i\|) \quad , \tag{4.10}$$

где $\{\phi(\|\boldsymbol{x}-\boldsymbol{x}_i\|), \quad i=1,2,\ldots,N\}$ - множество произвольных и обычно нелинейных функций, которые называются радиальными базисными функциями; норма используется обычно Евклидова. При этом выполняется следующее условие интерполяции:

$$F(\boldsymbol{x}_i)=d_i, \quad i=1,2,\ldots,N \quad , \tag{4.11}$$

где d_i - желаемый выходной сигнал при подаче на вход учебного вектора $\boldsymbol{x}_i$. Подставляя это условие в (4.10) можно получить систему уравнений для определения неизвестных весовых коэффициентов $\{w_i\}$;

$$\begin{bmatrix} \phi_{11} & \phi_{12} & \ldots & \phi_{1N} \\ \phi_{21} & \phi_{22} & \ldots & \phi_{2N} \\ \ldots & \ldots & \ldots & \ldots \\ \phi_{N1} & \phi_{N2} & \ldots & \phi_{NN} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \ldots \\ w_N \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \ldots \\ d_N \end{bmatrix} \quad , \tag{4.12}$$

где

$$\phi_{ji}=\phi(\|\boldsymbol{x}_j-\boldsymbol{x}_i\|), \quad i,j=1,2,\ldots,N \quad . \tag{4.13}$$

В матричном виде (4.12) можно записать как

$$\boldsymbol{\Phi w} = \boldsymbol{d} \quad . \tag{4.14}$$

Матрицу $\boldsymbol{\Phi}$ называют *матрицей интерполяции.* Если эта матрица является несингулярной, а для вычислительной реализации еще и хорошо обусловленной, то веса находятся просто:

$$\boldsymbol{w} = \boldsymbol{\Phi}^{-1}\boldsymbol{d} \quad . \tag{4.15}$$

Естественно возникает важный вопрос: как убедиться в несингулярности матрицы $\boldsymbol{\Phi}$? К счастью для большого класса радиальных базисных функций доказана следующая теорема Мичелли, (1985 г.).

Пусть $\{\boldsymbol{x}_i\}_{i=1}^{N}$ *- множество различных точек из* $\mathbb{R}^{m_0}$ *. Тогда матрица интерполяции* $\boldsymbol{\Phi}$ *размерности* $N\times N$ *с элементами* $\phi_{ji}=\phi(\|\boldsymbol{x}_j-\boldsymbol{x}_i\|)$ *является несингулярной.*

Теорема Мичелли охватывает большой класс радиальных базисных функции, среди которых наибольший интерес, с точки зрения сетей RBF, представляют следующие.

1. *Мультиквадратичная функция*:

$$\phi(r) = (r^2+c^2)^{1/2}, \quad c>0, r \in \mathbb{R} \,. \tag{4.16}$$

2. *Обратная мультиквадратичная функция*:

$$\phi = \frac{1}{(r^2+c^2)^{1/2}}, \quad c>0, r \in \mathbb{R} \,. \tag{4.17}$$

3. *Функция Гаусса*:

$$\phi = \exp\left(-\frac{r^2}{2\sigma^2}\right), \quad \sigma>0, r \in \mathbb{R} \,. \tag{4.18}$$

Для несингулярности матрицы $\boldsymbol{\Phi}$ требуется только то, чтобы все точки $\{\boldsymbol{x}_i\}_{i=1}^N$ были различными. Интересно то, что функции 2 и 3 являются *локализованными*, то есть $\phi(r) \to 0$ при $r \to 0$, а функция 1 наоборот неограниченно возрастает с ростом r, что не мешает ей тем не менее быть функцией пригодной для использования в сетях RBF.