К математической постановке задачи выбора архитектуры сети

Середа Яна

ННГУ им.Лобачевского
Нижний Новгород

The study was supported by the Ministry of Education and Science of Russia (Project No. 14.Y26.31.0022).

Дано:
Имеется набор данных $\{x_t\}_{t=1,..M}$
Он полагается выборкой из некоторой генеральной совокупности всех возможных исходов *X*, которая, в свою очередь, порождена некоторым генеративным процессом вида:

$$F(z_t) = x_t + noise$$

где noise - неустранимый аддитивный шум, встроенный в задачу
z_t вектор латентных (скрытых) переменных

Задача: - построить аппроксимацию для F

Сложности:
1) неизвестна размерность вектора z_t
2) неизвестно распределение вероятностей по разным значениям z_t
3) неизвестно, из какого параметрического семейства выбирать аппроксиматор ->выбирается избыточное кол-во параметров модели.

1. Zhang C. et al. Understanding deep learning requires rethinking generalization //arXiv preprint arXiv:1611.03530. – 2016.
2. Molchanov D., Ashukha A., Vetrov D. Variational dropout sparsifies deep neural networks //Proceedings of the 34th International Conference on Machine Learning-Volume 70. – JMLR. org, 2017. – C. 2498-2507.

1) сеть задается сразу вся - приходится надеяться, что угадали со структурой
(+ оверпарамеризация)

2) обучение сразу по данным всей задачи - “уравнивает” в значимости все эл-ты данных, что некорректно, учитывая неизбежные особенности выборки

необходимость в ручных способах увеличения выборки в “проблемных” местах

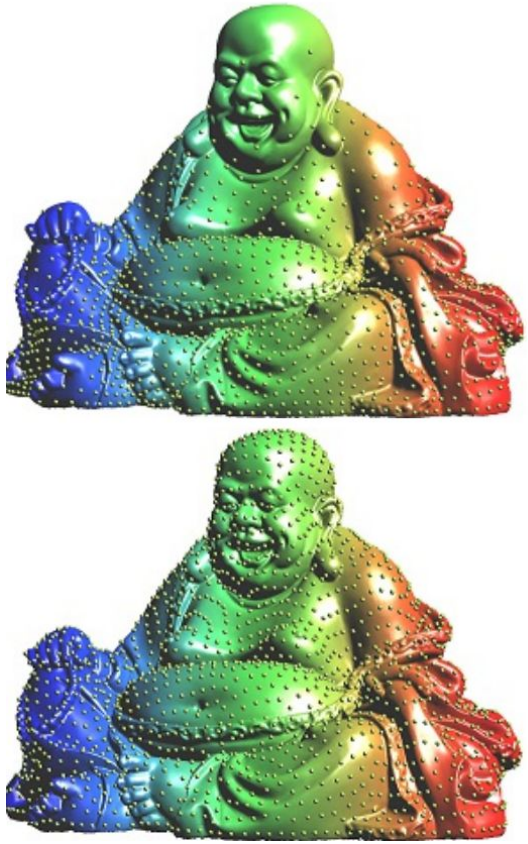


Рис.1.: Генеральная совокупность - многообразие размерности 2 ;

выборка всегда неравномерна!

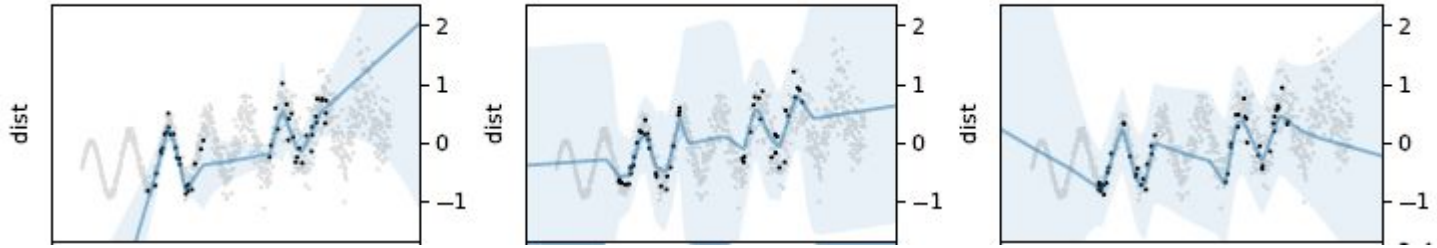


Figure 1: Predictive distributions on a low-dimensional active learning task. The predictive distributions are visualized as mean and two standard deviations shaded.

Hafner D. et al. Reliable uncertainty estimates in deep neural networks using noise contrastive priors //arXiv preprint arXiv:1807.09289. – 2018.

Obtaining reliable uncertainty estimates of neural network predictions is a long standing challenge. Bayesian neural networks have been proposed as a solution, but it remains open how to specify their prior.

Specifying priors is intuitive for small probabilistic models, where each variable often has a clear interpretation.

Bottlenecks of standard approach

Laplace's principle of indifference:
in the absence of any relevant evidence, agents should distribute their credence (or 'degrees of belief') equally among all the possible outcomes under consideration

The **principle of maximum entropy**:

the probability distribution which best represents the current state of knowledge is the one with largest entropy, in the context of precisely stated prior data (such as a proposition that expresses testable information).

Principle of maximum entropy: motivation

Physical problem: distribution of impurities in a crystal lattice

The mathematical problem: to find the $p(j|I)$ which will maximize the entropy

$$H = -\sum_{j=1}^{n} p(j|I) \log p(j|I)$$

subject to the constraints:

$$p(j|I) \geq 0$$

$$\sum_{j=1}^{n} p(j|I) = 1$$

natural constraints on the probability distribution

$$\sum_{j=1}^{n} p(j|I) \cos(kx_j) = 0.3.$$

prior information

Jaynes E. T. Prior probabilities //IEEE Transactions on systems science and cybernetics. – 1968. – Т. 4. – №. 3. – С. 227-241.

Решаем методом неопределенных множителей Лагранжа….

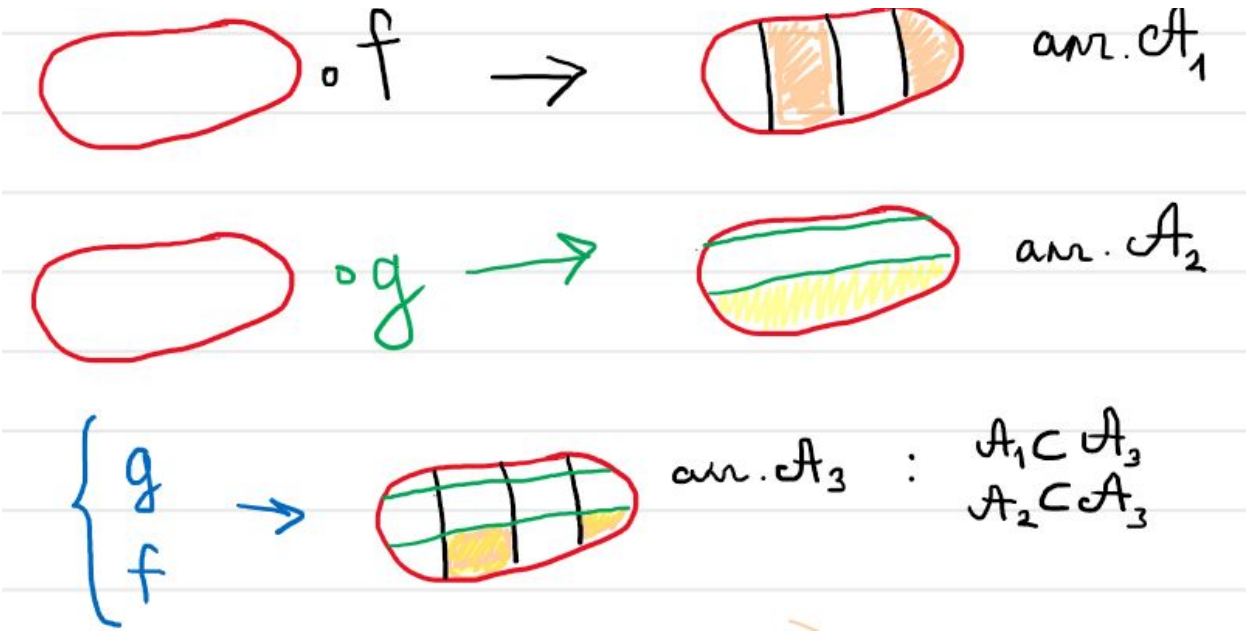
найдено экстремальное $p(j|I)$

можно использовать его как априорное для дальнейшего вывода (по теореме Байеса)

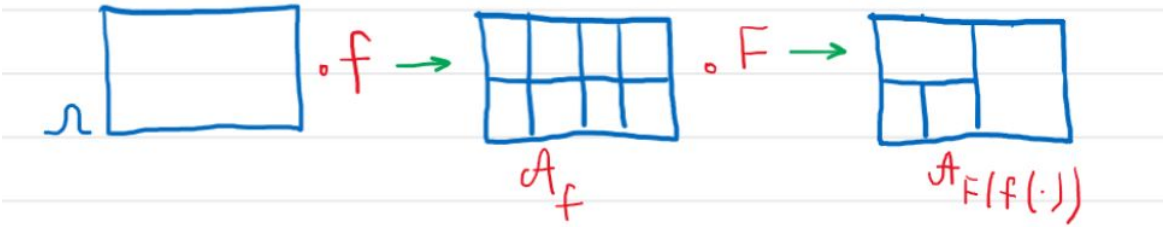
можем использовать его для оценки информативности набора ограничений в задаче оптимизации.

если оно не согласуется с новыми экспериментальными данными, имеется проблема с выбором информативных ограничений

The model can make reliable predictions only of those quantities for which it leads to a sharply peaked distribution.

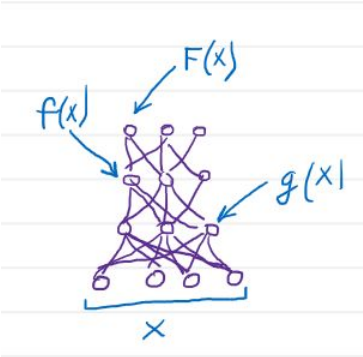


1. Каждое ограничение задает разбиение мн-ва исходов на классы эквивалентности
2. Одновременное выполнение нескольких ограничений измельчает разбиение



1. Композиция может только укрупнить разбиение

2. Чем выше нейрон находится в сети, тем крупнее разбиение на события, индуцируемое им в мн-ве элементарных исходов.



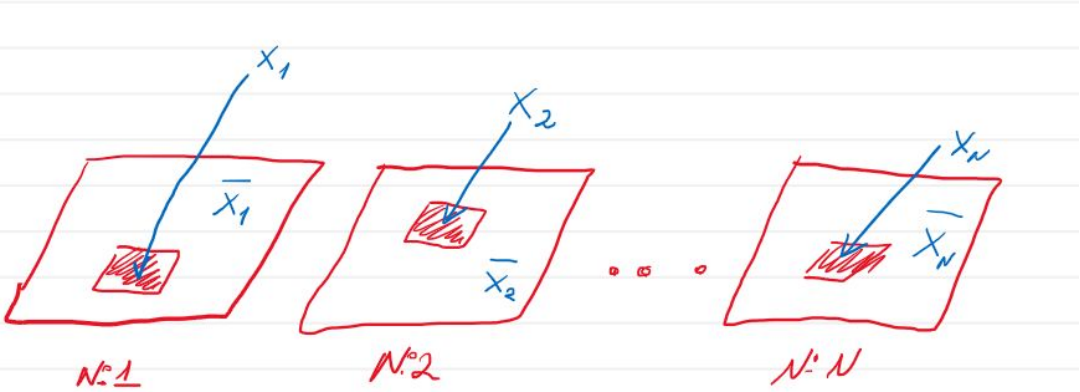
Principle of maximum entropy: let’s apply to DL? Step 2.

Изменим способ работы с датасетом.

Вместо обучения одновременно по всей выборке:

Сосредоточимся на маленькой однородной ее части (объединяющий признак выбран учителем)

Допустим, на выбранной группе картинок помечены регионы особого интереса.



Пример: так устроены обучающие материалы по медицине для людей.

$$\{X_1, X_2, \dots, X_N\}$$

-выборка значений интересного региона

Предположим, к данному моменту времени в сети имеется k нейронов

Запишем задачу поиска условного экстремума энтропии распределения для интересной области при условиях, даваемых имеющимися нейронами (теряющими информацию, об интеересной области)

Полученные множители Лагранжа покажут значимость ограничений

Отберем топ самых значимых → Найдем экстремальное распределение снова.

Если экстремальное распределение осталось с низкой энтропией, и выборка $\{X_1, X_2, \ldots, X_N\}$ имеет высокое правдоподобие в нем -> все ограничения в этом регионе учтены!

Ответ классификации тогда - набор ограничений-победителей!

Conclusion

При описанном подходе:

1) структура сети не задается априорно

Вместо этого её структура меняется в процессе поиска минимального набора достаточных статистик, который необходим для любых предсказаний в рамках моделируемой предметной области

2) сеть “знает”, когда она “хорошо” промоделировала сцену, и когда нет.