Deep Learning Book

Capítulo 3 -
Probabilidade e Teoria da Informação

Pedro Henrique Botecchi
pedrobotecchi@gmail.com

Teoria da Probabilidade

- Método matemático para representar afirmações incertas (incertezas);
- Meio de quantificar a incerteza e axiomas para derivar novas declarações incertas;
- Mostra-nos como a IA deve pensar e nos permite analisar o comportamento teórico de redes desenvolvidas.

Porque usar Probabilidade?

- Outras áreas da computação não se preocupam com dados não determinísticos.

- A área de Aprendizado de Máquina deve sempre lidar com quantidades incertas e muitas vezes estocásticas (não determinísticas).

Probabilidade

- Método para quantificar incertezas;

- É a representação de um grau de crença no qual 1 indica certeza absoluta de um evento ocorrer e 0 é a certeza de não ocorrência;

- A teoria da probabilidade fornece um conjunto de regras formais para determinar a probabilidade de uma proposição ser verdadeira, dada a probabilidade de outras proposições

Variáveis Aleatórias

- São variáveis que tomam valores diferentes para diferentes execuções de um mesmo evento;

- Podem ser Discretas ou Contínuas
  - Variáveis Aleatórias Discretas : Seu conjunto deve ser finito ou infinitamente enumerável;
  - Variáveis Aleatórias Contínuas : Seu conjunto é associado com valores reais.

Notação usada :

A variável aleatória é representada por uma letra maiúscula, enquanto que os possíveis assumidos são representados por letras minúsculas.

Distribuição de Probabilidade

- É a descrição do quão provável uma variável aleatória ou um conjunto de variáveis está de ser avaliada em um de seus possíveis valores;

- Variáveis Discretas e Função de Probabilidade de Massa:
  - Mapeia o estado de uma variável aleatória para a probabilidade dessa mesma variável assumir esse estado;

$$P(x) \qquad P(y)$$

$$P(X = x,\ Y = y) \qquad P(x, y)$$

$$P(X = x) = 1,\ \textit{evento certo};$$

$$P(X = x) = 0,\ \textit{evento impossível}$$

Propriedades

- Propriedades importantes :
  - O Domínio de P deve ser o conjunto de todos os possíveis estados de X
  - $\forall x \in X,\ 0 \leq P(x) \leq 1$
  - $\sum P(X = x) = 1$

Variáveis Contínuas e as Funções Densidade de Probabilidade

- Para variáveis contínuas nós descrevemos as distribuições de probabilidade utilizando as funções densidade de probabilidade (FDP) em vez das funções de probabilidade de massa;

- Propriedades:
  - O domínio de p deve ser o conjunto de todos os estados de x;
  - $\forall x \in X,\ 0 \leq P(x)$
  - $\int_{a}^{b} P(x)dx = 1$

Probabilidade Marginal

- Às vezes, quando temos a distribuição de probabilidade sobre um conjunto de variáveis e queremos saber a distribuição apenas sobre um subconjunto delas, ai utilizamos a probabilidade marginal :

$$\forall x \in x,\ P\ (x\ =\ x)\ =\sum_{y} P\ (x\ =\ x,\ y\ =\ y).$$

Probabilidade Condicional

- É utilizada quando queremos saber a probabilidade de um evento, dado que outro evento tenha acontecido

$$P\,(Y = y \mid X = x) = \frac{P\,(Y = y, X = x)}{P(X = x)}$$

Regra da Cadeia para Probabilidade Condicional

- Qualquer número de distribuições de probabilidade sobre um número n de variáveis aleatórias podem ser decompostos sobre uma distribuição condicional sobre uma variável:

$$P\,(x(1),\,\ldots,\,x(n)) \;=\; P\,(x(1)) \prod_{i=2}^{n} P\,(x(i)|\,x(1),\,\ldots,\,x(i-1))$$

Independência e Independência Condicional

- Duas variáveis aleatórias são independentes se suas distribuições de probabilidade podem sem expressas como um produto de dois fatores, um envolvendo X e outro envolvendo Y:

$$\forall x \in X,\ y \in Y,\ p(X = x,\ Y = y) = p(X = x)p(Y = y)$$

- E duas variáveis aleatórias X e Y são condicionalmente independentes dado uma variável aleatória Z se a distribuição de probabilidade condicional sobre X e Y pode ser fatorada da seguinte maneira para todo valor de Z:

$$\forall x \in X,\ y \in Y,\ z \in Z,\ p(X = x,\ Y = y \mid Z = z) = p(X = x \mid Z = z)p(Y = y \mid Z = z)$$

Esperança

- A Esperança de uma função f(x) com respeito a uma Distribuição de Probabilidade P(x) é dada por :

$$E_{X\sim P[f(x)]} = \sum_x P\ (x)f(x)$$

$$E_{X\sim P[f(x)]} = \int P\ (x)f(x)$$

Variância

- A variância fornece o quanto os valores de uma função de variável aleatória X varia conforme testamos diferentes valores de x de sua distribuição de probabilidade:

$$Var(f(x)) = E[f(x) - E[f(x)])^2]$$

A raiz quadrada da variância é conhecida como desvio padrão.

Covariância

- A Covariância mostra o senso de quanto dois valores estão linearmente relacionados a cada um , assim como a escala dessas duas variáveis:

$$Cov(f(x),\ g(y)) = E\left[(f(x) - E\left[f(x)\right])\,(g(y) - E\left[g(y)\right])\right]$$

Distribuições de Probabilidade mais comuns

- Distribuição de Bernuolli
- Distribuição Multiolli
- Distribuição de Gauss
- Distribuição de Laplace e Exponencial
- Distribuição Dirac e Distribuição Empírica

Distribuição de Bernuolli

- Distribuição sobre uma única variável aleatória que recebe valores binários. É controlada por uma único parâmetro no qual diz a probabilidade de uma variável aleatória apresentar o resultado esperado (= 1).

$$P(\mathrm{x} = 1) = \phi$$

$$P(\mathrm{x} = 0) = 1 - \phi$$

$$P(\mathrm{x} = x) = \phi^{x}(1 - \phi)^{1-x}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi$$

$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1 - \phi)$$

Distribuição de Multiolli

- A distribuição de Multiolli, ou distribuição categórica é uma distribuição sobre uma variável de k estados diferentes, sendo k finito

Distribuição de Gauss

- A distribuição mais comum usada sobre números reais em distribuições normais:

$$\mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

Exponencial e Distribuição de Laplace

- Em Deep Learning, normalmente é preferível ter uma distribuição de probabilidade com uma ponta em x=0, para isso podemos utilizar :

Exponential:

$$p(x;\lambda) = \lambda \mathbf{1}_{x \geq 0} \exp\left(-\lambda x\right)$$

Laplace:

$$\text{Laplace}(x;\mu,\gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x-\mu|}{\gamma}\right).$$

A Distribuição Dirac e a Distribuição Empírica

- Em alguns casos, desejamos especificar que toda a massa em uma distribuição de probabilidade se agrupe em torno de um único ponto. Isso pode ser feito definindo uma função densidade de probabilidade usando a função delta Dirac, δ (x) :

$$p(x) = \delta(x - \mu)$$

- Um uso comum da distribuição delta do Dirac é como componente de uma distribuição empírica:

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} \delta(\boldsymbol{x} - \boldsymbol{x}^{(i)})$$

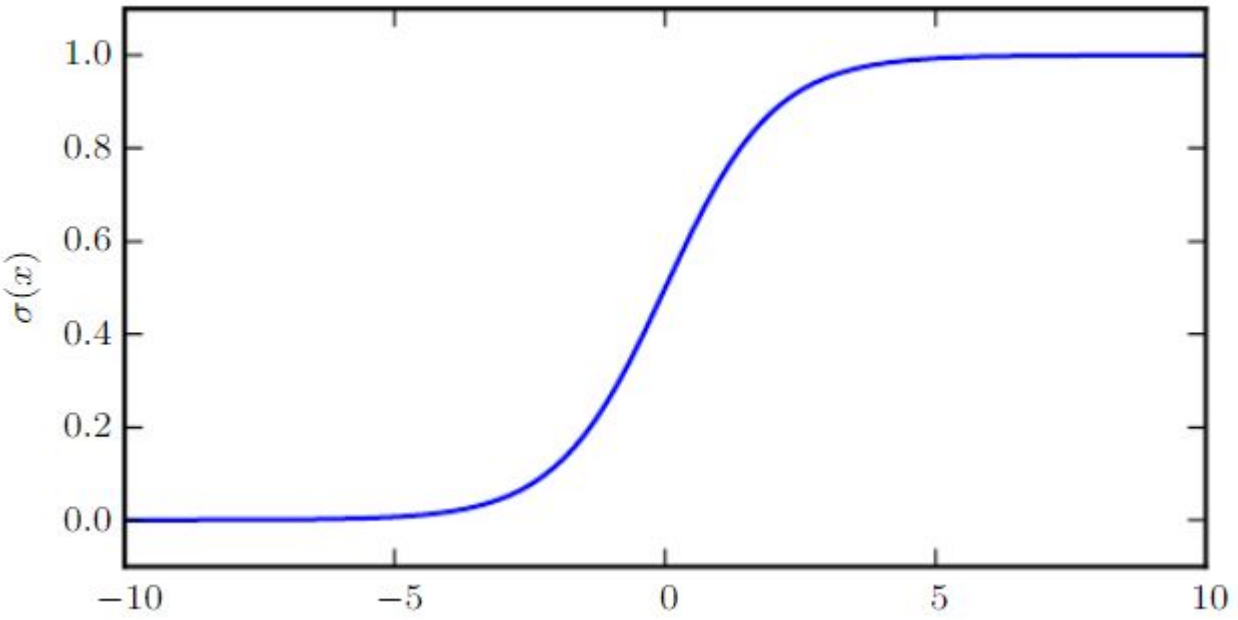
Propriedades Úteis de Funções Comuns

- Algumas funções surgem frequentemente ao trabalhar com distribuições de probabilidade, especialmente as distribuições de probabilidade usadas em modelos de aprendizado profundo:
  - Função Sigmoid
  - Função Softpuls

Função Sigmoid

- O sigmóide é comumente usado para produzir o parâmetro de uma distribuição Bernoulli porque seu int

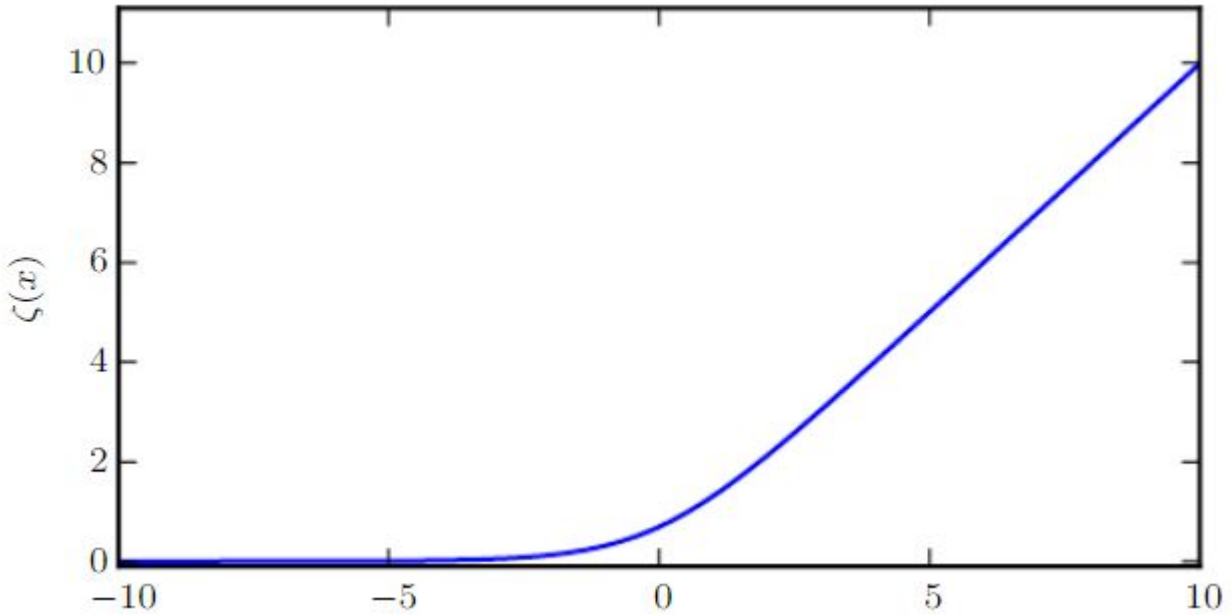
$$\sigma(x) = \frac{1}{1 + \exp(-x)}.$$



Softplus Function

- A função softplus pode ser útil para produzir o parâmetro β ou σ de uma distribuição normal porque seu intervalo é (0, ∞):

$$\zeta(x) = \log\left(1 + \exp(x)\right)$$



Propriedades Importantes

$$\sigma(x) = \frac{\exp(x)}{\exp(x) + \exp(0)}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$$

$$1 - \sigma(x) = \sigma(-x)$$

$$\log \sigma(x) = -\zeta(-x)$$

$$\frac{d}{dx}\zeta(x) = \sigma(x)$$

$$\forall x \in (0, 1),\ \sigma^{-1}(x) = \log\left(\frac{x}{1 - x}\right)$$

$$\forall x > 0,\ \zeta^{-1}(x) = \log\left(\exp(x) - 1\right)$$

$$\zeta(x) = \int_{-\infty}^{x} \sigma(y)dy$$

$$\zeta(x) - \zeta(-x) = x$$

Regra de Bayes

- Muitas vezes nos encontramos em uma situação em que conhecemos P (y | x) e precisamos conhecê-lo (x | y). Felizmente, se também conhecermos P (x), podemos calcular a quantidade desejada usando a regra de Bayes:

$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}$$

Teoria da Informação

- Gira em torno da quantificação da informação presente em um sinal;
- Originalmente inventado para estudar o envio de mensagens de alfabetos discretos através de um canal barulhento, como comunicação via transmissão de rádio;
- Nesse contexto, a teoria da informação diz como projetar códigos otimizados e calcular o comprimento esperado de mensagens amostradas de distribuições de probabilidade específicas usando vários esquemas de codificação.

Teoria da Informação

- A intuição básica por trás da teoria da informação é que aprender que um evento improvável ocorreu é mais informativo do que aprender que um evento provável ocorreu;
- "o sol nasceu esta manhã" x "houve um eclipse solar nesta manhã";

Quantificação da Informação

- Os eventos prováveis devem ter baixo conteúdo de informações e, no caso extremo, os eventos que são garantidos para acontecer não devem ter nenhum conteúdo de informações;

- Eventos menos prováveis devem ter maior conteúdo informativo;

- Eventos independentes devem ter informações aditivas. Por exemplo, descobrir que uma moeda lançada surgiu como cara duas vezes deve transmitir duas vezes mais informações do que descobrir que uma moeda lançada surgiu como cara uma vez.

Auto-Informação

- Para satisfazer todas essas três propriedades, definimos as informações de um evento X = x a ser:

$$I(x) = -\log P(x)$$

- Nossa definição de I (x) é, portanto, escrita em unidades de nats. Um nat é a quantidade de informações obtidas pela observação de um evento de probabilidade 1/e.

Entropia de Shannon

- Podemos quantificar a quantidade de incerteza em uma distribuição de probabilidade inteira usando a entropia de Shannon :

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]$$

- Em outras palavras, a entropia de Shannon de uma distribuição é a quantidade esperada de informações em um evento extraído dessa distribuição

Divergência de Kullback-Leibler

- Se tivermos duas distribuições de probabilidade separadas P (x) e Q (x) sobre a mesma variável aleatória, podemos medir a diferença dessas duas distribuições usando a divergência Kullback-Leibler (KL):

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}[\log P(x) - \log Q(x)]$$

Cross Entropy

- Uma quantidade que está intimamente relacionada à divergência KL é a entropia cruzada, que é semelhante à divergência KL, mas sem o termo à esquerda:

$$H(P, Q) = -\mathbb{E}_{\mathbf{x} \sim P} \log Q(x)$$

Modelos Probabilísticos Estruturados

- Os algoritmos de aprendizado de máquina geralmente envolvem distribuições de probabilidade em um número muito grande de variáveis aleatórias. Frequentemente, essas distribuições de probabilidade envolvem interações diretas entre relativamente poucas variáveis.
- Ineficiência ao usar 1 função

Modelos Probabilísticos Estruturados

For example, suppose we have three random variables: a, b and c. Suppose that a influences the value of b, and b influences the value of c, but that a and c are independent given b. We can represent the probability distribution over all three variables as a product of probability distributions over two variables:

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{b}). \tag{3.52}$$

OBRIGADO!

Dúvidas?