Deep Learning

AlejandroMllo

This document serves as a very brief summary of the topics covered in each chapter of the book Deep Learning [1].

Disclaimer: This document is completely extracted from [1], the author does not attribute any ownership over the material.

3 Probability and Information Theory

- Possible sources of uncertainty:
  1. Inherent stochasticity in the system being modeled.
  2. Incomplete observability.
  3. Incomplete modeling.
- Probability can be seen as the extension of logic to deal with uncertainty.
- Random Variable: variable x (discrete or continuous) that can take on different values randomly.
- Probability Distribution (PD): description of how likely a random variable or a set of random variables is to take on each of its possible states.
- Probability Mass Function (PMF): PD over discrete variables. A PMF, P, maps from a state of a random variable to the probability of that random variable taking on that state. A PMF acting on many variables at the same time is known as a joint probability distribution: $P(\mathrm{x} = x, \mathrm{y} = y)$.
- Probability Density Function (PDF): describes PD of continuous random variables. A PDF, $p(x)$, gives the probability of landing inside an infinitesimal region, not at a specific state.
- Marginal PD: it is the PD of a subset of a set of variables of which we know the PD.
- Conditional Probability: the probability of some event $\mathrm{x} = x$, given that some other event $\mathrm{y} = y$ has happened.
$$P(\mathtt{y} = y \mid \mathtt{x} = x) = \frac{P(\mathtt{y} = y\ ,\ \mathtt{x} = x)}{P(\mathtt{x} = x)},\ \textit{where } P(\mathtt{x} = x) > 0.$$
- Chain Rule of Conditional Probabilities: any joint PD over many random variables may be decomposed into conditional distributions over only one variable:
$$P(\mathtt{x}^{(1)}, \ldots, \mathtt{x}^{(n)}) = P(\mathtt{x}^{(1)}) \prod_{i=2}^{n} P(\mathtt{x}^{(i)} \mid \mathtt{x}^{(1)}, \ldots, \mathtt{x}^{(i-1)}).$$
- Two random variables x and y are independent ($\mathrm{x} \perp \mathrm{y}$) if:
$$\forall x \in \mathtt{x}, y \in \mathtt{y}, p(\mathtt{x} = x, \mathtt{y} = y) = p(\mathtt{x} = x)p(\mathtt{y} = y)$$
- Two random variables x and y are conditionally independent given a random variable z ($\mathrm{x} \perp \mathtt{y} | z$) if:
$$\forall x \in \mathtt{x}, y \in \mathtt{y}, z \in \mathtt{z}, p(\mathtt{x} = x, \mathtt{y} = y \mid \mathtt{z} = z) = p(\mathtt{x} = x \mid \mathtt{z} = z)p(\mathtt{y} = y \mid \mathtt{z} = z)$$
- The expectation ($\mathbb{E}_{x \sim P}[f(x)]$), or expected value, of some function $f(x)$ w.r.t. a PD $P(\mathrm{x})$ is the average, or mean value, that f takes on when x is drawn from P.
$$\mathbb{E}_{x \sim P}[f(x)] = \sum_{x} P(x)f(x),\ \textit{for discrete variables.}$$
$$\mathbb{E}_{x \sim p}[f(x)] = \int p(x)f(x)dx,\ \textit{for continuous variables.}$$
- The variance gives a measure of how much the values of a function of a random variable x vary as we sample different values of x from its PD. The square root of the variance is the standard deviation.
$$\mathtt{Var}(f(X)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

- The covariance gives some sense of how much two values are linearly related to each other, as well as the scale of these variables. Two variables have zero covariance if they are not linearly dependent.

$$\texttt{Cov}(f(X), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])]$$

- The correlation normalizes the contribution of each variable in order to measure only how much the variables are related.
- The covariance matrix of a random vector $\mathbf{x} \in \mathbb{R}^n$ is an $n \times n$ matrix, such that $\mathrm{Cov}(\mathrm{x})_{i,j} = \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j)$. The diagonal elements of the covariance give the variance: $\mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_i) = \mathrm{Var}(\mathrm{x}_i)$.
- The Bernoulli Distribution is a distribution over a single binary random variable. It is controlled by a single parameter $\phi \in [0, 1]$, which gives the probability of the random variable being equal to 1. $P(\mathtt{x} = x) = \phi^x(1-\phi)^{1-x}$, $\mathbb{E}_{\mathtt{x}}[\mathtt{x}] = \phi$, $\texttt{Var}_{\mathtt{x}}(\mathtt{x}) = \phi(1-\phi)$.
- The multinoulli, or categorical, distribution is a distribution over a single discrete variable with k different states, where k is finite. It is parametrized by a vector $\boldsymbol{p} \in [0, 1]^{k-1}$, where p_i gives the probability of the i-th state. The final, k-th state's probability is given by 1 - $1^\top \boldsymbol{p}$.
- Gaussian, normal, distribution: distribution over the real numbers. The parameters $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$ control the normal distribution.

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

  To evaluate the PDF efficiently, parametrize the distribution with $\beta \in (0, \infty)$, to control the precision, or inverse variance:

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

  Multivariate normal distribution: the normal distribution generalized to $\mathbb{R}^n$. It may be parametrized with a positive definite symmetric matrix $\mathbf{\Sigma}$:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

  where $\boldsymbol{\mu}$ is a vector with the mean of the distribution, and $\mathbf{\Sigma}$ gives the covariance matrix.
  Use a precision matrix $\boldsymbol{\beta}$, to evaluate several times for different values of the parameters:

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

- Exponential distribution: PD with a sharp point at $x = 0$: $p(x; \lambda) = \lambda \mathbf{1}_{x \geq 0} \exp(-\lambda x)$. It uses $\mathbf{1}_{x \geq 0}$ to assign probability zero to all negative values of x.
- Laplace distribution: PD that places a sharp peak of probability mass at an arbitrary point μ: $\mathrm{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp(-\frac{|x-\mu|}{\gamma})$.
- Dirac delta function: makes possible to specify that all the mass in a PD clusters around a point.
- Mixture distribution: PD made up of several component distributions.
- Latent variable: random variable that is not possible to observe directly.
- Logistic sigmoid:

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

- Softplus function: (smoothed version of $x^+ = \max(0, x)$)

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

- Bayes' Rule:

$$P(\mathtt{x}|\mathtt{y}) = \frac{P(\mathtt{x})P(\mathtt{y}|\mathtt{x})}{P(\mathtt{y})}$$

  It is possible to compute $P(\mathrm{y})$ as: $P(\mathtt{y}) = \sum_x P(\mathtt{y}|\mathtt{x})P(\mathtt{x})$
- Information theory: quantifies how much information is present in a signal.
  - Basic intuition: learning that an unlikely event has occurred is more informative than learning that a likely event has occurred. Likely events have low (or zero) information content. Less likely events have higher information content. Independents events have additive information content.

- Self-information of an event $\mathrm{x} = x$: $I(x) = -\log P(x) \quad [nats]$. One nat is the amount of information gained by observing an event of probability $\frac{1}{e}$. (The book uses log as the natural logarithm).
- Shannon entropy: quantifies the amount of uncertainty in an entire PD. $H(\mathtt{x}) = \mathbb{E}_{x\sim P}[I(x)]$.
- To measure how different two distributions over the same random variable are, use the Kullback-Leibler (KL) divergence: $D_{KL}(P||Q) = \mathbb{E}_{x\sim P}[\log P(x) - \log Q(x)]$.
- Cross-entropy: $H(P,Q) = -\mathbb{E}_{x\sim P}\log Q(x)$.

- A Structured probabilistic model, or graphical model, represents the factorization of a PD with a graph (represents the PD as factors). Each node in the graph is a random variable, and an edge connecting two random variables means that the PD is able to represent direct interaction between those two random variables.
  - Directed models represent factorization into conditional PD.

$$p(\mathtt{x}) = \prod_i p(\mathtt{x}_i|PaG(x_i)),$$

    where PaG(x_i) represents the parents of x_i.
  - Undirected models represent factorizations into a set of functions. Any set of nodes connected to each other is called a clique. Each clique $C(i)$ is associated with a factor $\phi^{(i)}(C^{(i)})$.

$$p(\mathtt{x}) = \frac{1}{Z}\prod_i \phi^{(i)}(C^{(i)}),$$

    where Z is a normalizing constant defined to be the sum or integral over all states of the product of the ϕ functions.

References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.