Probability and Information Theory

Lecture slides for Chapter 3 of *Deep Learning*
www.deeplearningbook.org
Ian Goodfellow
2016-09-26
Probability Mass Function

- The domain of $P$ must be the set of all possible states of $x$.
- $\forall x \in x, 0 \leq P(x) \leq 1$. An impossible event has probability 0 and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring.
- $\sum_{x \in x} P(x) = 1$. We refer to this property as being normalized. Without this property, we could obtain probabilities greater than one by computing the probability of one of many events occurring.

Example: uniform distribution: $P(x = x_i) = \frac{1}{k}$
Probability Density Function

- The domain of \( p \) must be the set of all possible states of \( x \).
- \( \forall x \in x, p(x) \geq 0 \). Note that we do not require \( p(x) \leq 1 \).
- \( \int p(x)dx = 1 \).

Example: uniform distribution: \( u(x; a, b) = \frac{1}{b-a} \).
Computing Marginal Probability with the Sum Rule

\[ \forall x \in \mathcal{X}, P(x = x) = \sum_{y} P(x = x, y = y). \quad (3.3) \]

\[ p(x) = \int p(x, y) dy. \quad (3.4) \]
Conditional Probability

\[ P(y = y \mid x = x) = \frac{P(y = y, x = x)}{P(x = x)}. \] (3.5)
Chain Rule of Probability

\[ P(x^{(1)}, \ldots, x^{(n)}) = P(x^{(1)}) \prod_{i=2}^{n} P(x^{(i)} \mid x^{(1)}, \ldots, x^{(i-1)}). \]  

(3.6)
Independence

\[ \forall x \in x, y \in y, \quad p(x = x, y = y) = p(x = x)p(y = y). \] (3.7)
Conditional Independence

\[ \forall x \in x, y \in y, z \in z, \ p(x = x, y = y \mid z = z) = p(x = x \mid z = z)p(y = y \mid z = z). \]  
(3.8)
**Expectation**

\[
\mathbb{E}_{x \sim P}[f(x)] = \sum_{x} P(x) f(x), \quad (3.9)
\]

\[
\mathbb{E}_{x \sim p}[f(x)] = \int p(x) f(x) \, dx. \quad (3.10)
\]

linearity of expectations:

\[
\mathbb{E}_x[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_x[f(x)] + \beta \mathbb{E}_x[g(x)], \quad (3.11)
\]

(Goodfellow 2016)
### Variance and Covariance

\[ \text{Var}(f(x)) = \mathbb{E} \left[ (f(x) - \mathbb{E}[f(x)])^2 \right]. \quad (3.12) \]

\[ \text{Cov}(f(x), g(y)) = \mathbb{E} \left[ (f(x) - \mathbb{E}[f(x)]) (g(y) - \mathbb{E}[g(y)]) \right]. \quad (3.13) \]

**Covariance matrix:**

\[ \text{Cov}(x)_{i,j} = \text{Cov}(x_i, x_j). \quad (3.14) \]
Bernoulli Distribution

\[ P(x = 1) = \phi \]  
\[ P(x = 0) = 1 - \phi \]  
\[ P(x = x) = \phi^x (1 - \phi)^{1-x} \]  
\[ \mathbb{E}_x[x] = \phi \]  
\[ \text{Var}_x(x) = \phi(1 - \phi) \]
Gaussian Distribution

Parametrized by variance:

\[ \mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \]  
(3.21)

Parametrized by precision:

\[ \mathcal{N}(x; \mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{1}{2\beta}(x - \mu)^2\right). \]  
(3.22)
Gaussian Distribution

Figure 3.1: The normal distribution \( N(x; \mu, \sigma) \) exhibits a classic "bell curve" shape, with the \( x \) coordinate of its central peak given by \( \mu \), and the width of its peak controlled by \( \sigma \). In this example, we depict the standard normal distribution, with \( \mu = 0 \) and \( \sigma = 1 \).

First, many distributions we wish to model are truly close to being normal distributions. The central limit theorem shows that the sum of many independent random variables is approximately normally distributed. This means that in practice, many complicated systems can be modeled successfully as normally distributed noise, even if the system can be decomposed into parts with more structured behavior.

Second, out of all possible probability distributions with the same variance, the normal distribution encodes the maximum amount of uncertainty over the real numbers. We can thus think of the normal distribution as being the one that inserts the least amount of prior knowledge into a model. Fully developing and justifying this idea requires more mathematical tools, and is postponed to section 19.4.2.

The normal distribution generalizes to \( \mathbb{R}^n \), in which case it is known as the multivariate normal distribution. It may be parametrized with a positive definite symmetric matrix \( \Sigma \):

\[
N(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right). 
\]
Multivariate Gaussian

Parametrized by covariance matrix:

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

(3.23)

Parametrized by precision matrix:

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

(3.24)
More Distributions

Exponential:

\[ p(x; \lambda) = \lambda 1_{x \geq 0} \exp (-\lambda x). \]  \hspace{1cm} (3.25)

Laplace:

\[ \text{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp \left(-\frac{|x - \mu|}{\gamma}\right). \]  \hspace{1cm} (3.26)

Dirac:

\[ p(x) = \delta(x - \mu). \]  \hspace{1cm} (3.27)
Empirical Distribution

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

(3.28)
Mixture Distributions

\[ P(x) = \sum_i P(c = i)P(x \mid c = i) \]  \hspace{1cm} (3.29)

Gaussian mixture
with three
components

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{gaussian_mixture.png}
\caption{Gaussian mixture with three components}
\end{figure}
Logistic Sigmoid

Figure 3.3: The logistic sigmoid function.

Commonly used to parametrize Bernoulli distributions
Softplus Function

Figure 3.4: The softplus function.
Bayes’ Rule

\[ P(x \mid y) = \frac{P(x)P(y \mid x)}{P(y)}. \quad (3.42) \]
Change of Variables

\[ p_x(x) = p_y(g(x)) \left| \det \left( \frac{\partial g(x)}{\partial x} \right) \right|. \]  

(3.47)
CHAPTER 3. PROBABILITY AND INFORMATION THEORY

Specifically, "morning" is very informative. To be unnecessary to send, but a message saying "there was a solar eclipse this morning" would be much information as finding out that a tossed coin has come up as heads twice should convey twice as much information. Likely events should have low information content, and in the extreme case, an event that is guaranteed to occur. Information theory is a branch of applied mathematics that revolves around probability distributions or quantify similarity between probability distributions. The basic intuition behind information theory is that learning that an unlikely event has occurred is very informative. To be unnecessary to send, but a message saying "there was a solar eclipse this morning" would be much information as finding out that a tossed coin has come up as heads twice should convey twice as much information. Likely events should have low information content, and in the extreme case, an event that is guaranteed to occur.

Information theory is a branch of applied mathematics that revolves around probability distributions or quantify similarity between probability distributions. The basic intuition behind information theory is that learning that an unlikely event has occurred is very informative.

In order to satisfy all three of these properties, we define the self-information of an outcome $x$ as:

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

This definition captures the intuitive idea that the information content of an event depends on its probability: the lower the probability, the higher the information content.

The expected information content of a distribution $P(x)$ is the entropy of the distribution:

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

This gives a lower bound on the expected number of bits needed to encode an event drawn from the distribution. It gives the expected amount of information in an event drawn from that distribution.

The KL divergence is a measure of the distance between two probability distributions $P(x)$ and $Q(x)$:

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

This measure is not symmetric: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, and it is not a true distance measure because it is not necessarily finite. However, it is a fundamental concept in information theory and machine learning.

Information theory is a branch of applied mathematics that revolves around probability distributions or quantify similarity between probability distributions. The basic intuition behind information theory is that learning that an unlikely event has occurred is very informative.

In order to satisfy all three of these properties, we define the self-information of an outcome $x$ as:

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

This definition captures the intuitive idea that the information content of an event depends on its probability: the lower the probability, the higher the information content.

The expected information content of a distribution $P(x)$ is the entropy of the distribution:

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

This gives a lower bound on the expected number of bits needed to encode an event drawn from the distribution. It gives the expected amount of information in an event drawn from that distribution.

The KL divergence is a measure of the distance between two probability distributions $P(x)$ and $Q(x)$:

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

This measure is not symmetric: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, and it is not a true distance measure because it is not necessarily finite. However, it is a fundamental concept in information theory and machine learning.
Figure 3.5: This plot shows how distributions that are closer to deterministic have low Shannon entropy while distributions that are close to uniform have high Shannon entropy. On the horizontal axis, we plot $p$, the probability of a binary random variable being equal to 1. The entropy is given by $(p \log p + (1 - p) \log (1 - p))$. When $p$ is near 0, the distribution is nearly deterministic, because the random variable is nearly always 0. When $p$ is near 1, the distribution is nearly deterministic, because the random variable is nearly always 1. When $p = 0.5$, the entropy is maximal, because the distribution is uniform over the two outcomes. Asymmetry means that there are important consequences to the choice of whether to use $D_{KL}(P \| Q)$ or $D_{KL}(Q \| P)$. See figure 3.6 for more detail. A quantity that is closely related to the KL divergence is the cross-entropy $H(P, Q) = H(P) + D_{KL}(P \| Q)$, which is similar to the KL divergence but lacking the term on the left: $H(P, Q) = \sum_{x \in \mathcal{X}} P(x) \log Q(x)$. Minimizing the cross-entropy with respect to $Q$ is equivalent to minimizing the KL divergence, because $Q$ does not participate in the omitted term.
The KL Divergence is Asymmetric

\[ q^* = \arg\min_q D_{KL}(p\|q) \]

\[ q^* = \arg\min_q D_{KL}(q\|p) \]

Figure 3.6
\[ p(a, b, c, d, e) = p(a)p(b \mid a)p(c \mid a, b)p(d \mid b)p(e \mid c). \]
Figure 3.8: An undirected graphical model over random variables \(a, b, c, d\) and \(e\). This graph corresponds to probability distributions that can be factored as

\[
p(a, b, c, d, e) = \frac{1}{Z} \phi^{(1)}(a, b, c)\phi^{(2)}(b, d)\phi^{(3)}(c, e).
\]  

(3.56)