Neural Networks

Lecture 2: Probability and Information theory
<table>
<thead>
<tr>
<th>Lesson</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
</tr>
<tr>
<td>2</td>
<td>Probability and Information Theory</td>
</tr>
<tr>
<td>3</td>
<td>Basics of Machine Learning</td>
</tr>
<tr>
<td>4</td>
<td>Deep Feed-forward networks</td>
</tr>
<tr>
<td>5</td>
<td>Back-propagation</td>
</tr>
<tr>
<td>6</td>
<td>Optimization and regularization</td>
</tr>
<tr>
<td>7</td>
<td>Convolutional Networks</td>
</tr>
<tr>
<td>8</td>
<td>Sequence Modeling: recurrent networks</td>
</tr>
<tr>
<td>9</td>
<td>Applications</td>
</tr>
<tr>
<td>10</td>
<td>Software and Practical methods</td>
</tr>
<tr>
<td>11</td>
<td>Autoencoders &amp; GANs</td>
</tr>
<tr>
<td>12</td>
<td>Attention and Transformers</td>
</tr>
<tr>
<td>13</td>
<td>Deep Reinforcement Learning</td>
</tr>
<tr>
<td>14</td>
<td>Deep Learning and the Brain</td>
</tr>
<tr>
<td>15</td>
<td>Conclusions and Future Perspective</td>
</tr>
<tr>
<td>16</td>
<td>Project presentations</td>
</tr>
</tbody>
</table>
Logic is exact but brittle
Sources of uncertainty

Inherent stochasticity

Incomplete observability

Incomplete modeling
Keep it simple

“Most birds fly”

“Birds fly, except the very young birds that have not yet learned to fly, sick or injured birds that have lost their ability to fly, flightless species of birds including cassowary, ostrich and kiwi…”
Real data is noisy and uncertain

**Kinds of Noise**

**Attributes**
- Att 1: 0.25, red, positive
- Att 2: 0.25, red, negative
- Att 1: 0.99, green, negative
- Att 1: 1.02, green, positive
- Att 1: 2.05, ?, negative
- Att 1: =, green, positive

**Class**
- positive
- negative

**Class Noise**
- Contradictory examples
- Mislabeled examples

**Attribute Noise**
- Erroneous values
- Missing values
- Don’t care values
There are two types of people in this world:

1) Those who can extrapolate from incomplete data.
Probability theory

Framework to represent uncertain statements and derive new uncertain statements

\[ p(A) = 0 \rightarrow p(A) = 1 \]
Probability theory

Framework to represent uncertain statements and derive new uncertain statements

Information theory

Quantify amount of uncertainty in a probability distribution
Probabilistic regression
We can think of a neural network as representing a function $f(x, w)$

The output of this function are not direct predictions of the value $y$

$f(x, w)$ provides the **parameters** for a distribution over $y$

$$p(y|x:w) = \text{Gaussian}(y; \text{mean}=f(x, w), \text{variance}=1)$$
Learning objectives

• to review basic concepts of probability and information theory

• to understand the use of neural networks to model probability functions
Primer in Probability

Mini-primer in Information theory
Primer in Probability

Mini-primer in Information theory
Random variables

**Random variable**: a variable that can take on different values randomly
A **probability distribution**: description of how likely a random variable (or a set of RV) is to take on each of its possible states

**Discrete**
- $X$ can take finite or countably infinite values $x_1, x_2, \ldots$

**Continuous**
- $X$ can take continuous values $\mathbb{R}^n$
Discrete

**Probability Mass Function (PMF):** list of probabilities of random variable $X$ taking each particular value

- 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: 

$$P(x = x_i) = \frac{1}{k}$$
Continuous

**Probability Density Function (PDF):** considers the probability of $X$ falling in a small interval with volume $dx$ as $p(x)dx$

- 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: $\ u(x; a, b) = \frac{1}{b-a}$.
Expectation

Average or mean value of \( f(x) \) when we sample \( x \) from \( P \):

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

(3.9)

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

(3.10)

Linearity of expectations:

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

(3.11)

(Goodfellow 2016)
Variance and Covariance

Variance quantifies how much sample vary from their mean:

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

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

Covariance matrix:

\[
\text{Cov}(\mathbf{x})_{i,j} = \text{Cov}(x_i, x_j).
\] (3.14)
# Most popular statistics

<table>
<thead>
<tr>
<th>name</th>
<th>notation</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>$&lt; X &gt;$, $\mu_X$</td>
<td>$E[X]$</td>
</tr>
<tr>
<td>variance</td>
<td>$Var[X]$, $\sigma^2_X$</td>
<td>$E[(X - E[X])^2] = E[X^2] - E[X]^2$</td>
</tr>
<tr>
<td>correlation</td>
<td>$Cor[X, Y]$</td>
<td>$\frac{Cov[X,Y]}{E[X]E[Y]}$</td>
</tr>
</tbody>
</table>
Joint and marginal probability

If there are 2 or more random variables, say $X$ and $Y$, we can consider their **joint** probability of taking a particular pair of values, $P(X=x, Y=y)$. 

<table>
<thead>
<tr>
<th>Event</th>
<th>$B_1$</th>
<th>$B_2$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>$P(A_1 \text{ and } B_1)$</td>
<td>$P(A_1 \text{ and } B_2)$</td>
<td>$P(A_1)$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$P(A_2 \text{ and } B_1)$</td>
<td>$P(A_2 \text{ and } B_2)$</td>
<td>$P(A_2)$</td>
</tr>
<tr>
<td>Total</td>
<td>$P(B_1)$</td>
<td>$P(B_2)$</td>
<td>1</td>
</tr>
</tbody>
</table>

**Joint Probabilities**

**Marginal (Simple) Probabilities**
Joint and marginal probability

We know the probability distribution over a set of variables and we want to know the probability distribution over just a subset of them. The pd over the subset is called the **marginal**
Sum rule

\[ \forall x \in x, P(x = x) = \sum_y P(x = x, y = y). \]  

(3.3)

\[ p(x) = \int p(x, y) dy. \]  

(3.4)
CHAPTER 3. PROBABILITY AND INFORMATION THEORY

The name "marginal probability" comes from the process of computing marginal probabilities on paper. When the values of $P(x, y)$ are written in a grid with different values of $x$ in rows and different values of $y$ in columns, it is natural to sum across a row of the grid, then write $P(x)$ in the margin of the paper just to the right of the row.

For continuous variables, we need to use integration instead of summation:

$$p(x) = \int p(x, y) \, dy.$$  

(3.4)

3.5 Conditional Probability

In many cases, we are interested in the probability of some event, given that some other event has happened. This is called a conditional probability. We denote the conditional probability that $y = y$ given $x = x$ as $P(y = y \mid x = x)$. This conditional probability can be computed with the formula

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

(3.5)

The conditional probability is only defined when $P(x = x) > 0$. We cannot compute the conditional probability conditioned on an event that never happens.

It is important not to confuse conditional probability with computing what would happen if some action were undertaken. The conditional probability that a person is from Germany given that they speak German is quite high, but if a randomly selected person is taught to speak German, their country of origin does not change. Computing the consequences of an action is called making an intervention query. Intervention queries are the domain of causal modeling, which we do not explore in this book.

3.6 The Chain Rule of Conditional Probabilities

Any joint probability distribution over many random variables may be decomposed into conditional distributions over only one variable:

$$P(x_1, \ldots, x_n) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_1, x_2) \cdot \ldots \cdot P(x_n \mid x_1, x_2, \ldots, x_{n-1}).$$

(3.6)

This observation is known as the chain rule or product rule of probability. It follows immediately from the definition of conditional probability in equation 3.5.
Questionnaire

<table>
<thead>
<tr>
<th></th>
<th>no wind</th>
<th>some wind</th>
<th>strong wind</th>
<th>storm</th>
</tr>
</thead>
<tbody>
<tr>
<td>no rain</td>
<td>0.1</td>
<td>0.2</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>light rain</td>
<td>0.05</td>
<td>0.1</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>heavy rain</td>
<td>0.05</td>
<td>0.1</td>
<td>0.1</td>
<td>0.05</td>
</tr>
</tbody>
</table>

\[
P(\text{no wind})
\]

\[
P(\text{light rain})
\]

\[
P(\text{no wind}|\text{light rain})
\]

\[
P(\text{light rain}|\text{no wind})
\]
Questionnaire

<table>
<thead>
<tr>
<th></th>
<th>no wind</th>
<th>some wind</th>
<th>strong wind</th>
<th>storm</th>
</tr>
</thead>
<tbody>
<tr>
<td>no rain</td>
<td>0.1</td>
<td>0.2</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>light rain</td>
<td>0.05</td>
<td>0.1</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>heavy rain</td>
<td>0.05</td>
<td>0.1</td>
<td>0.1</td>
<td>0.05</td>
</tr>
</tbody>
</table>

\[ P(\text{no wind}) = 0.1 + 0.05 + 0.05 = 0.2 \]
\[ P(\text{light rain}) = 0.05 + 0.1 + 0.15 + 0.04 = 0.34 \]
\[ P(\text{no wind} | \text{light rain}) = \frac{0.05}{0.34} = 0.147 \]
\[ P(\text{light rain} | \text{no wind}) = \frac{0.05}{0.2} = 0.25 \]
CHAPTER 3. PROBABILITY AND INFORMATION THEORY

The name “marginal probability” comes from the process of computing marginal probabilities on paper. When the values of $P(x, y)$ are written in a grid with different values of $x$ in rows and different values of $y$ in columns, it is natural to sum across a row of the grid, then write $P(x)$ in the margin of the paper just to the right of the row.

For continuous variables, we need to use integration instead of summation:

$$p(x) = \int p(x, y) \, dy.$$ 

(3.4)

3.5 Conditional Probability

In many cases, we are interested in the probability of some event, given that some other event has happened. This is called a conditional probability. We denote the conditional probability that $y = y$ given $x = x$ as $P(y = y| x = x)$. This conditional probability can be computed with the formula

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

(3.5)

The conditional probability is only defined when $P(x = x) > 0$. We cannot compute the conditional probability conditioned on an event that never happens.

It is important not to confuse conditional probability with computing what would happen if some action were undertaken. The conditional probability that a person is from Germany given that they speak German is quite high, but if a randomly selected person is taught to speak German, their country of origin does not change. Computing the consequences of an action is called making an intervention query. Intervention queries are the domain of causal modeling, which we do not explore in this book.

3.6 The Chain Rule of Conditional Probabilities

Any joint probability distribution over many random variables may be decomposed into conditional distributions over only one variable:

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

(3.6)
Bayes’ Rule

\[ P(x \mid y) = \frac{P(x)P(y \mid x)}{P(y)} . \] (3.42)

\[ P(\text{hypothesis} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis})P(\text{hypothesis})}{P(\text{data})} \]
CHAPTER 3. PROBABILITY AND INFORMATION THEORY

For example, applying the definition twice, we get

\[ P(a, b, c) = P(a | b, c) P(b, c) = P(b | c) P(c) \]

\[ P(a, b, c) = P(a | b, c) P(b | c) P(c) \]

3.7 Independence and Conditional Independence

Two random variables \( x \) and \( y \) are independent if their probability distribution can be expressed as a product of two factors, one involving only \( x \) and one involving only \( y \):

\[
p(x = x, y = y) = p(x = x)p(y = y). \tag{3.7}
\]

\[
P(X | Y) = P(X), \quad P(Y | X) = P(Y).
\]

Uncorrelated variables:

\[ E[XY] = E[X]E[Y]. \]

Independent variables are uncorrelated but the reverse is not true!

Independence and correlation
CHAPTER 3. PROBABILITY AND INFORMATION THEORY

For example, applying the definition twice, we get

\[ P(a, b, c) = P(a | b, c) P(b, c) = P(b | c) P(c) \]

\[ P(a, b, c) = P(a | b, c) P(b | c) P(c) \]

3.7 Independence and Conditional Independence

Two random variables \( x \) and \( y \) are independent if their probability distribution can be expressed as a product of two factors, one involving only \( x \) and one involving only \( y \):

\[ p(x = x, y = y) = p(x = x) p(y = y) \] (3.7)

Two random variables \( x \) and \( y \) are conditionally independent given a random variable \( z \) if the conditional probability distribution over \( x \) and \( y \) factorizes in this way for every value of \( z \):

\[ p(x = x, y = y | z = z) = p(x = x | z = z) p(y = y | z = z) \] (3.8)

We can denote independence and conditional independence with compact notation:

\( x \perp y \) means that \( x \) and \( y \) are independent, while \( x \perp y \mid z \) means that \( x \) and \( y \) are conditionally independent given \( z \).

3.8 Expectation, Variance and Covariance

The expectation or expected value of some function \( f(x) \) with respect to a probability distribution \( P(x) \) is the average or mean value that \( f \) takes on when \( x \) is drawn from \( P \).

For discrete variables this can be computed with a summation:

\[ E_{x \sim P}[f(x)] = \sum_x f(x) P(x) \] (3.9)

while for continuous variables, it is computed with an integral:

\[ E_{x \sim p}[f(x)] = \int p(x) f(x) \, dx \] (3.10)

Conditional independence

\[ \forall x \in x, y \in y, z \in z, \ p(x = x, y = y | z = z) = p(x = x | z = z) p(y = y | z = z). \] (3.8)

Myopia and ambient lighting at night

- Night light
- Myopic children

**Night light**

**Myopic children**

1 year later... (also in *Nature*):

**Night light**

**Myopic parents**

**Myopic children**

---

**Myopia and ambient lighting at night**

Graham E. Quinn*,†, Chai H. Shin*, Maureen G. Maguire†, Richard A. Stone†
*Department of Ophthalmology, Children’s Hospital of Philadelphia, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA
†Department of Ophthalmology, Scheie Eye Institute, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA
e-mail: stone2@mail.med.upenn.edu

Karla Zadnik*, Lisa A. Jones*, Brett C. Irvin*, Robert N. KleinStein†, Ruth E. Manny‡, Julie A. Shin§, Donald O. Mutti*, for the CLEERE Study Group
*College of Optometry, Ohio State University, Columbus, Ohio 43210-1240, USA
†School of Optometry, University of Alabama, Birmingham, Alabama 35294-0010, USA
‡College of Optometry, University of Houston, Houston, Texas 77204-6052, USA
§Southern California College of Optometry, Fullerton, California 92831, USA
e-mail: zadnik.4@osu.edu

**Myopia and ambient night-time lighting**

J. Gwiazda, E. Ong, R. Held, F. Thorn
New England College of Optometry, 424 Beacon Street, Boston, Massachusetts 02115, USA
Bernoulli Distribution

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

51% same side as started
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)
CHAPTER 3. PROBABILITY AND INFORMATION THEORY

2.0

1.5

1.0

0.5

0.0

-0.5

-1.0

-1.5

-2.0

p(x)

Maximum at $x = \mu$

Inflection points at $x = \mu \pm \sigma$

Figure 3.1: The normal distribution $N(x; \mu, \sigma^2)$ 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 $\Sigma$:

$$N(x; \mu, \Sigma) = (2\pi)^{n/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu)\right).$$  (3.23)
Multivariate Gaussian

Parametrized by covariance matrix:

\[ N(x; \mu, \Sigma) = \sqrt{\frac{1}{(2\pi)^n \det(\Sigma)}} \exp \left( -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right). \] (3.23)

Parametrized by precision matrix:

\[ N(x; \mu, \beta^{-1}) = \sqrt{\frac{\det(\beta)}{(2\pi)^n}} \exp \left( -\frac{1}{2} (x - \mu)^\top \beta (x - \mu) \right). \] (3.24)
## Popular distributions

<table>
<thead>
<tr>
<th>name</th>
<th>definition</th>
<th>range</th>
<th>mean</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial</td>
<td>( \binom{N}{x} \alpha^x (1 - \alpha)^{N-x} )</td>
<td>( x = 0, 1, ..., N )</td>
<td>( N\alpha )</td>
<td>( N\alpha(1 - \alpha) )</td>
</tr>
<tr>
<td>Poisson</td>
<td>( \frac{1}{x!} \alpha^x e^{-\alpha} )</td>
<td>( x = 0, 1, 2, ... )</td>
<td>( \alpha )</td>
<td>( \alpha )</td>
</tr>
<tr>
<td>Gaussian or normal</td>
<td>( \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} )</td>
<td>( x \in \mathbb{R} )</td>
<td>( \mu )</td>
<td>( \sigma^2 )</td>
</tr>
<tr>
<td>Gamma</td>
<td>( \frac{1}{\Gamma(a)} b^a x^{a-1} e^{-bx} ) (\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx ) (\Gamma(a) = a! ) if ( a ) is an integer)</td>
<td>( x \geq 0 )</td>
<td>( \frac{a}{b} )</td>
<td>( \frac{a}{b^2} )</td>
</tr>
</tbody>
</table>
Primer in Probability

Mini-primer in Information theory
Measuring information

Surprise:

\[ \log \frac{1}{P(X = x)} = - \log P(X = x) \]
Measuring information

Surprise:

$$\log \frac{1}{P(X = x)} = - \log P(X = x)$$

Additivity for independent events:

$$\log \frac{1}{P(x, y)} = \log \frac{1}{P(x)P(y)} = \log \frac{1}{P(x)} + \log \frac{1}{P(y)}$$
Entropy

**Shanon entropy**: expected amount of information (surprise) in an event drawn from that distribution

\[ H(X) = E[-\log P(X)] = \sum_X -P(X) \log P(X) \]

Number of bits needed to encode symbols drawn from the distribution \( P \)
Entropy of a Bernoulli Variable

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 $(1 - p) \log(1 - p) + p \log p$. 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.

A quantity that is closely related to the KL divergence is the cross-entropy $H(P, Q) = H(P) + D_{KL}(P \parallel Q)$, which is similar to the KL divergence but lacking the term on the left: $H(P, Q) = \mathbb{E}_{x \sim P} \log Q(x)$. (3.51)

Minimizing the cross-entropy with respect to $Q$ is equivalent to minimizing the KL divergence, because $Q$ does not participate in the omitted term.

When computing many of these quantities, it is common to encounter expressions of the form $0 \log 0$. By convention, in the context of information theory, we treat these expressions as $\lim_{x \to 0} x \log x = 0$. 3.14 Structured Probabilistic Models

Machine learning algorithms often involve probability distributions over a very large number of random variables. Often, these probability distributions involve direct interactions between relatively few variables. Using a single function
KL divergence

**Kullback-Leibler divergence**: measures how different 2 separate distributions $P$ and $Q$ over the same random variable $X$

$$\log \frac{1}{Q(x)} - \log \frac{1}{P(x)} = \log \frac{P(x)}{Q(x)}$$

$$D(P; Q) = E_{P(x)} \left[ \log \frac{P(x)}{Q(x)} \right] = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

$$D(P, Q) \neq D(Q, P)$$
Cross-entropy

Closely related to KL divergence:

\[ H(p, q) = - \sum_x p(x) \log q(x) \]

\[ H(p, q) = \mathbb{E}_p[-\log q] = H(p) + D_{\text{KL}}(p\|q) \]
Cross-entropy

Closely related to KL divergence:

\[
H(p, q) = - \sum_x p(x) \log q(x)
\]

\[
H(p, q) = E_p[- \log q] = H(p) + D_{KL}(p || q)
\]

Example (Bernouilli):

\[
p \in \{y, 1 - y\} \quad q \in \{\hat{y}, 1 - \hat{y}\}
\]

\[
H(p, q) = - \sum_i p_i \log q_i = - y \log \hat{y} - (1 - y) \log (1 - \hat{y})
\]
Cross-entropy

Closely related to KL divergence:

\[ H(p, q) = - \sum_x p(x) \log q(x) \]

\[ H(p, q) = E_p[-\log q] = H(p) + D_{KL}(p || q) \]

Example (Bernouilli):

\[ p \in \{y, 1-y\} \quad q \in \{\hat{y}, 1-\hat{y}\} \]

\[ H(p, q) = - \sum_i p_i \log q_i = -y \log \hat{y} -(1-y) \log (1-\hat{y}) \]

Used as a loss function in ML (between P_data and P_model)
Application to ML
The network

\[ y = f^{(2)}(h; w, b) \]

\[ h = f^{(1)}(x; W, c) \]

\[ f(x; W, c, w, b) = f^{(2)}(f^{(1)}(x)) \]
The output of this function are not direct predictions of the value $y$; $f(x, W)$ provides the **parameters** for a distribution over $y$

$$f(x; W, c, w, b) = f^{(2)}(f^{(1)}(x)) \quad \rightarrow \quad p_{\text{model}}(y \mid x) = \mathcal{N}(y; f(x; \theta), I)$$
We want to learn statistical properties of underlying distribution from analysis of data. Select a statistical model...

Underlying distribution $p_{\text{data}}(\mathbf{x})$

$\mathbf{X} = \{ \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \}$

$\hat{p}_{\text{data}}$

$f(\mathbf{x}; \mathbf{W}, \mathbf{c}, \mathbf{w}, b) = f^{(2)}(f^{(1)}(\mathbf{x}))$ $p_{\text{model}}(\mathbf{x}; \theta)$

$\theta$ $\theta_{\text{ML}} = \arg \max_{\theta} p_{\text{model}}(\mathbf{X}; \theta)$
Maximum likelihood

\[ \theta_{ML} = \arg \max_{\theta} p_{\text{model}}(X; \theta) \]

\[ = \arg \max_{\theta} \prod_{i=1}^{m} p_{\text{model}}(x^{(i)}; \theta) \quad \text{(iid)} \]

\[ \theta_{ML} = \arg \max_{\theta} \sum_{i=1}^{m} \log p_{\text{model}}(x^{(i)}; \theta) \quad \text{(monotonicity of log)} \]

\[ \theta_{ML} = \arg \max_{\theta} \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(x; \theta) \quad \text{(def. of Expectation)} \]
Maximize $\mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(x; \theta)$
Maximize \[ \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(x; \theta) \]

KL divergence

\[ D_{\text{KL}} (\hat{p}_{\text{data}} \parallel p_{\text{model}}) = \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \left[ \log \hat{p}_{\text{data}}(x) - \log p_{\text{model}}(x) \right] \]
Maximize $\mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(x; \theta)$

KL divergence

$$D_{KL}(\hat{p}_{\text{data}} \parallel p_{\text{model}}) = \mathbb{E}_{x \sim \hat{p}_{\text{data}}} [\log \hat{p}_{\text{data}}(x) - \log p_{\text{model}}(x)]$$

Cross-entropy

$$H(p, q) = - \sum_{x} p(x) \log q(x)$$
$$H(p, q) = \mathbb{E}_{p} [-\log q] = H(p) + D_{KL}(p \parallel q)$$
Maximize \[ \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(x; \theta) \]

KL divergence

\[ D_{\text{KL}} (\hat{p}_{\text{data}} \parallel p_{\text{model}}) = \mathbb{E}_{x \sim \hat{p}_{\text{data}}} \left[ \log \hat{p}_{\text{data}}(x) - \log p_{\text{model}}(x) \right] \]

Cross-entropy

\[ H(p, q) = -\sum_x p(x) \log q(x) \]

\[ H(p, q) = \mathbb{E}_p [-\log q] = H(p) + D_{\text{KL}}(p\parallel q) \]

Used as a loss function in ML (between \( P_{\text{data}} \) and \( P_{\text{model}} \))
Maximum likelihood

Unsupervised
(Unconditional)

\[ \theta_{ML} = \arg \max_{\theta} p_{\text{model}}(X; \theta) \]

\[ \theta_{ML} = \arg \max_{\theta} \sum_{i=1}^{m} \log p_{\text{model}}(x^{(i)}; \theta) \]

Supervised
(Conditional)

\[ \theta_{ML} = \arg \max_{\theta} P(Y \mid X; \theta) \]

\[ \theta_{ML} = \arg \max_{\theta} \sum_{i=1}^{m} \log P(y^{(i)} \mid x^{(i)}; \theta) \]
Summary
Real data is noisy and uncertain

Information Sources

Attributes

<table>
<thead>
<tr>
<th>Att 1</th>
<th>Att 2</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>red</td>
<td>positive</td>
</tr>
<tr>
<td>0.25</td>
<td>red</td>
<td>negative</td>
</tr>
<tr>
<td>0.99</td>
<td>green</td>
<td>negative</td>
</tr>
<tr>
<td>1.02</td>
<td>green</td>
<td>positive</td>
</tr>
<tr>
<td>2.05</td>
<td>?</td>
<td>negative</td>
</tr>
<tr>
<td>=</td>
<td>green</td>
<td>positive</td>
</tr>
</tbody>
</table>

Class Noise

- Contradictory examples
- Mislabeled examples

Attribute Noise

- Erroneous values
- Missing values
- Don’t care values

Kinds of Noise

Att. Noise  Class Noise
We can think of a neural network as representing a function \( f(x,w) \)
The output of this function are not direct predictions of the value \( y \)
\( f(x,w) \) provides the parameters for a distribution over \( y \)
\[
p(y|x:w) = \text{Gaussian}(y; f(x,w), 1)
\]
If there are 2 or more random variables, say $X$ and $Y$, we can consider their joint probability of taking a particular pair of values, $P(X=x, Y=y)$. 

### Joint and marginal probability

<table>
<thead>
<tr>
<th>Event</th>
<th>$B_1$</th>
<th>$B_2$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>$P(A_1 \text{ and } B_1)$</td>
<td>$P(A_1 \text{ and } B_2)$</td>
<td>$P(A_1)$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$P(A_2 \text{ and } B_1)$</td>
<td>$P(A_2 \text{ and } B_2)$</td>
<td>$P(A_2)$</td>
</tr>
<tr>
<td>Total</td>
<td>$P(B_1)$</td>
<td>$P(B_2)$</td>
<td>1</td>
</tr>
</tbody>
</table>

**Joint Probabilities**

**Marginal (Simple) Probabilities**
Maximize $\mathbb{E}_{x \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(x; \theta)$

KL divergence

$$D_{KL}(\hat{p}_{\text{data}} \parallel p_{\text{model}}) = \mathbb{E}_{x \sim \hat{p}_{\text{data}}} [\log \hat{p}_{\text{data}}(x) - \log p_{\text{model}}(x)]$$

Cross-entropy

$$H(p, q) = -\sum_x p(x) \log q(x)$$

$$H(p, q) = \mathbb{E}_p [-\log q] = H(p) + D_{KL}(p \parallel q)$$

Used as a loss function in ML (between $P_{\text{data}}$ and $P_{\text{model}}$)
<table>
<thead>
<tr>
<th>Lesson</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
</tr>
<tr>
<td>2</td>
<td>Probability and Information Theory</td>
</tr>
<tr>
<td>3</td>
<td>Basics of Machine Learning</td>
</tr>
<tr>
<td>4</td>
<td>Deep Feed-forward networks</td>
</tr>
<tr>
<td>5</td>
<td>Back-propagation</td>
</tr>
<tr>
<td>6</td>
<td>Optimization and regularization</td>
</tr>
<tr>
<td>7</td>
<td>Convolutional Networks</td>
</tr>
<tr>
<td>8</td>
<td>Sequence Modeling: recurrent networks</td>
</tr>
<tr>
<td>9</td>
<td>Applications</td>
</tr>
<tr>
<td>10</td>
<td>Software and Practical methods</td>
</tr>
<tr>
<td>11</td>
<td>Autoencoders &amp; GANs</td>
</tr>
<tr>
<td>12</td>
<td>Project Development</td>
</tr>
<tr>
<td>13</td>
<td>Deep Reinforcement Learning</td>
</tr>
<tr>
<td>14</td>
<td>Deep Learning and the Brain</td>
</tr>
<tr>
<td>15</td>
<td>Conclusions and Future Perspective</td>
</tr>
<tr>
<td>16</td>
<td>Project presentations</td>
</tr>
</tbody>
</table>
**Mutual information:** measures how far 2 random variables $X$ and $Y$ are from being independent

\[
I(X; Y) = D_{KL}(P(X, Y) \| P(X)P(Y))
\]

\[
D(P; Q) = E_P(x) \left[ \log \frac{P(x)}{Q(x)} \right] = \sum_X P(x) \log \frac{P(x)}{Q(x)}
\]

\[
I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)
\]