CS-567 Machine Learning

Nazar Khan

PUCIT

Lecture 9
Information Theory

Information Theory

- Amount of additional information $\propto$ degree of surprise.
- If a highly unlikely event occurs, you gain a lot of new information.
- If an almost certain event occurs, you gain not much new information.
- So information $\propto \frac{1}{\text{probability}}$

Information Theory

- For unrelated events x and y
  - Information from both events should equal information from x plus information from y.
  - $p(x, y) = p(x)p(y)$
- From these two relationships, it can be shown that information must be given by the logarithm function.

$$
\begin{aligned}
h(x, y) &= -\log(p(x, y)) \\
&= -\log(p(x)p(y)) \\
&= -\log(p(x)) - \log(p(y)) \\
h(x) &= -\log(p(x))
\end{aligned}
$$

  where $h(x)$ denotes the information given by x.
- For base 2 log, units of information $h(x)$ are 'bits'.
- For natural log, units of information $h(x)$ are 'nats' (1 nat$= \ln 2$ bits).

Information Theory
Entropy

- If information given by random variable x is given by a function $h(x) = -\log(p(x))$, then <u>expected information from r.v x</u> is

$$H[x] = E[h(x)] = -\sum \log(p(x))p(x)$$

- Also called the **entropy** of random variable x.
- Entropy is just a fancy name for expected information contained in a random variable.

Information Theory
Entropy

- To transmit a r.v x with 8 *equally likely* states, we need 3 bits ($= \log_2 8$).
- Entropy $H[x] = -\sum \frac{1}{8} \log_2 \frac{1}{8} = 3$ bits.
- For non-uniform probabilities, entropy is reduced.
- **Entropy quantifies order/disorder.**
- Entropy is a lower-bound on the number of bits needed to transmit the state of a random variable.

Information Theory

Entropy

- For a *discrete* r.v X with pdf p, entropy is

$$H[p] = -\sum_i p(x_i) \ln p(x_i) \tag{1}$$

- Sharply peaked distribution $\implies$ low entropy.
- Evenly spread distribution $\implies$ high entropy.
- Is the entropy non-negative?
- What is its minimum value?
- When does the minimum value occur?

Information Theory

Finding the Maximum Entropy Distribution – Discrete Case

- How can we find the *discrete* distribution $p(x)$ that maximises the entropy $H[p]$?
- Since p must add up to 1, this a constrained maximisation problem.
- The Lagrangian function is

$$\tilde{H} = -\sum_i p(x_i) \ln p(x_i) + \lambda \left(\sum_i p(x_i) - 1 \right)$$

- The maximum is given by the stationary point of $\tilde{H}$.
- Why is it the maximum?

Information Theory
Entropy

- For a *continuous* r.v X with pdf p, we define **differential entropy** as

$$H[p] = -\int p(x) \ln p(x) dx$$

- For multivariate $\mathbf{x}$

$$H[p] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}$$

Information Theory

Finding the Maximum Entropy Distribution – Discrete Case

- How can we find the *continuous* distribution $p(x)$ that maximises the entropy $H[p]$?
- The maximum entropy discrete distribution was the **uniform** distribution.
- The maximum differential entropy continuous distribution is the **Gaussian** distribution (Excercise 1.34 in Bishop's book).

Information Theory
Entropy

- Differential entropy of the Gaussian is

$$H[x] = \frac{1}{2}\{1 + \ln(2\pi\sigma^2)\}$$

- Proportional to σ^2. Entropy increases as more values become probable.
- Can also be negative (for $\sigma^2 < \frac{1}{2\pi e}$).

Information Theory
Conditional Entropy

- Let $p(\mathbf{x}, \mathbf{y})$ be a joint distribution.
- Given $\mathbf{x}$, additional information needed to specify $\mathbf{y}$ is the conditional information $-\ln(p(\mathbf{y}|\mathbf{x}))$.
- So expected conditional information is

$$H[\mathbf{y}|\mathbf{x}] = -\int\int p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{y} d\mathbf{x}$$

- Also called the **conditional entropy** of $\mathbf{y}$ given $\mathbf{x}$.
- Satisfies $H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}]$. Information needed to specify $\mathbf{x}$ and $\mathbf{y}$ equals information for $\mathbf{x}$ alone plus *additional* information needed to specify $\mathbf{y}$ given $\mathbf{x}$.

Information Theory
Relative entropy

- Let r.v. $\mathbf{x}$ have a true distribution $p(\mathbf{x})$ and let our estimate of this distribution be $q(\mathbf{x})$.
- Average information required to specify x when its information content is determined using $p(\mathbf{x})$ is given by the entropy

$$H[p] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) \tag{2}$$

- Average information required to specify x when its information content is determined using $q(\mathbf{x})$ is given by

$$\tilde{H}[q] = -\int p(\mathbf{x}) \ln q(\mathbf{x}) \tag{3}$$

Information Theory
Relative entropy

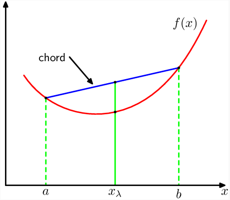
- Average *additional* information required to specify x when $q(\mathbf{x})$ is used instead of $p(\mathbf{x})$ is given by $\tilde{H}[q] - H[p] = \left(-\int p(\mathbf{x}) \ln q(\mathbf{x})\right) - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\right)$.
- This is known as the **relative entropy**, or **Kullback-Leibler (KL) divergence**.

$$
\begin{aligned}
KL(p||q) &= \left(-\int p(\mathbf{x}) \ln q(\mathbf{x})\right) d\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\right) d\mathbf{x} \\
&= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} d\mathbf{x}
\end{aligned}
$$

- $KL(p||q) \neq KL(q||p)$.
- $KL(p||q) \geq 0$ with equality for $p = q$.

Convex Functions

- A function $f(x)$ is **convex** if every chord lies on or above the function.
- Any value of x in the interval a to b can be parameterised as $\lambda a + (1 - \lambda)b$ where $0 \leq \lambda \leq 1$.
- The corresponding point on the chord can be parameterised as $\lambda f(a) + (1 - \lambda)f(b)$.
- The corresponding point on the function can be parameterised as $f(\lambda a + (1 - \lambda)b)$.



Convex Functions

- Convexity implies points on chord lie on or above points on function. That is

$$f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b)$$

- Convexity is equivalent to positive second derivative everywhere.
- If function and chord are equal only for $\lambda = 0$ and $\lambda = 1$, then the function is called **strictly convex**.
- The inverse property (every chord lies on or below the function) is called **concavity**.
- If $f(x)$ is convex, then $-f(x)$ will be concave.

Jensen's Inequality

- Every convex function $f(x)$ satisfies the so-called **Jensen's inequality**

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i)$$

where $\lambda_i \geq 0$ and $\sum_{i=1}^{M} \lambda_i = 1$ for any set of points $(x_1, \ldots, x_M)$.

- Interpreting the λ_i as probabilities $p(x_i)$, Jensen's inequality can be formulated for *discrete random variables* as

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$

- For *continuous random variables*, Jensen's inequality becomes

$$f\left(\int \mathbf{x} p(\mathbf{x} d\mathbf{x}\right) \leq \int f(\mathbf{x})\, p(\mathbf{x} d\mathbf{x}$$

KL-divergence

- Using Jensen's inequality

$$KL(p||q) = -\int p(\mathbf{x}) \underbrace{\ln\left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\}}_{\text{concave}} d\mathbf{x} \geq -\underbrace{\ln \underbrace{\int q(\mathbf{x}) d\mathbf{x}}_{=1}}_{=0}$$

where the equality holds only when $p(\mathbf{x}) = q(\mathbf{x})\ \forall \mathbf{x}$ (because $-\ln x$ is strictly convex).

- Since $KL(p||q) \geq 0$ and $KL(p||p) = 0$, KL-divergence can be interpreted as a **measure of dissimilarity** between distributions $p(\mathbf{x})$ and $q(\mathbf{x})$.

Relation between data compression and density estimation

- Optimal compression requires the true density.
- For estimated density, KL-divergence gives **average, additional information** required by **transmitting via estimated density** instead of true density.

Density Estimation via KL-divergence

- Suppose we have finite data points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ drawn from an *unknown* distribution $p(\mathbf{x})$.
- We want to approximate $p(\mathbf{x})$ by some parametric distribution $q(\mathbf{x}|\boldsymbol{\theta})$.
- We can do this by finding $\boldsymbol{\theta}$ that minimizes $KL(p||q)$. **But** p **is unknown.**
- However, $KL(p||q)$ is an *expectation w.r.t* $p(\mathbf{x})$ and can be approximated by the ordinary average for large N (law of large numbers). So

$$KL(p||q) = -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x})} \right\} d\mathbf{x} \qquad (4)$$

$$\approx \frac{1}{N} \sum_{n=1}^{N} \{ -\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}) \}$$

Density Estimation via KL-divergence

- Minimizing w.r.t $\boldsymbol{\theta}$ is equivalent to minimizing $\sum_{n=1}^{N} - \ln q(\mathbf{x}_n|\boldsymbol{\theta})$ which is the **negative log-likelihood of data under** $q(\mathbf{x}|\boldsymbol{\theta})$.
- So *minimizing KL-divergence is equivalent to maximising likelihood (ML estimation)*.

Mutual Information

- Given 2 random variables $\mathbf{x}$ and $\mathbf{y}$, can we find *how independent* they are?
- If they are independent then $p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y})$. So $KL(p(\mathbf{x}, \mathbf{y})||p(\mathbf{x})p(\mathbf{y})) = 0$.
- Therefore, $KL(p(\mathbf{x}, \mathbf{y})||p(\mathbf{x})p(\mathbf{y}))$ is a measure of *how independent* $\mathbf{x}$ and $\mathbf{y}$ are.
- Also called the **mutual information** $I[\mathbf{x}, \mathbf{y}]$ between variables $\mathbf{x}$ and $\mathbf{y}$.

$$
\begin{aligned}
I[\mathbf{x}, \mathbf{y}] &= KL(p(\mathbf{x}, \mathbf{y})||p(\mathbf{x})p(\mathbf{y})) \qquad (5) \\
&= -\int\int p(\mathbf{x}, \mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})}\right) d\mathbf{x} d\mathbf{y}
\end{aligned}
$$

- $I[\mathbf{x}, \mathbf{y}] \geq 0$ with equality iff $\mathbf{x}$ and $\mathbf{y}$ are independent.

Mutual Information

- Using the sum and product rules

$$I[\mathbf{x},\mathbf{y}] = \underbrace{H[\mathbf{x}]}_{\substack{\text{avg. info. needed} \\ \text{to transmit } \mathbf{x}}} - \underbrace{H[\mathbf{x}|\mathbf{y}]}_{\substack{\text{avg. info. needed} \\ \text{to transmit } \mathbf{x} \\ \text{knowing state of } \mathbf{y}}}$$

$$= \underbrace{H[\mathbf{y}]}_{\substack{\text{avg. info. needed} \\ \text{to transmit } \mathbf{y}}} - \underbrace{H[\mathbf{y}|\mathbf{x}]}_{\substack{\text{avg. info. needed} \\ \text{to transmit } \mathbf{y} \\ \text{knowing state of } \mathbf{x}}}$$

- Mutual information captures
  - Information about **x** that is contained in **y**.
  - Information about **y** that is contained in **x**.
  - Reduction in uncertainty of one variable when the other is known.