Introduction to Statistics
Session 2

Grzegorz Chrupała

Saarland University

October 18, 2012

Outline

1. Random variables and information theory

2. Discrete probability distributions

Random variables

- Function $X : \Omega \to \mathbb{R}^n$ (typically $n = 1$)
- It may be more convenient to work with real number than directly with events

Random variables

- Function $X : \Omega \to \mathbb{R}^n$ (typically $n = 1$)
- It may be more convenient to work with real number than directly with events
- Coin toss: $X : \{H, T\} \to \{0, 1\}$

Random variables

- Function $X : \Omega \to \mathbb{R}^n$ (typically $n = 1$)
- It may be more convenient to work with real number than directly with events
- Coin toss: $X : \{H, T\} \to \{0, 1\}$
- Sum of two dice throws: $\{1..6\}^2 \to \{2..12\}$

Random variables

- Function $X : \Omega \to \mathbb{R}^n$ (typically $n = 1$)
- It may be more convenient to work with real number than directly with events
- Coin toss: $X : \{H, T\} \to \{0, 1\}$
- Sum of two dice throws: $\{1..6\}^2 \to \{2..12\}$
- Probability mass function:

$$\mathrm{p}(x) = P(X = x) = P(A) \text{ where } A = \{\omega \in \Omega : X(\omega) = x\}$$

Expectation

- Expectation is a mean (weighted average) of a random variable

$$E(X) = \sum_x \mathrm{p}(x) \cdot x$$

- Example: rolling a dice:

$$E(X) =$$

Expectation

- Expectation is a mean (weighted average) of a random variable

$$E(X) = \sum_x \mathrm{p}(x) \cdot x$$

- Example: rolling a dice:

$$E(X) = \sum_{x=1}^{6} \mathrm{p}(x)x = \sum_{x=1}^{6} \frac{x}{6} = 3.5$$

Expectation

- Expectation is a mean (weighted average) of a random variable

$$E(X) = \sum_x \mathrm{p}(x) \cdot x$$

- Example: rolling a dice:

$$E(X) = \sum_{x=1}^{6} \mathrm{p}(x)x = \sum_{x=1}^{6} \frac{x}{6} = 3.5$$

- A function $g(X)$ defines new random variable. In this case:

$$E(g(X)) = \sum_x \mathrm{p}(x)g(x)$$

Example?

Expectation

- Expectation is a mean (weighted average) of a random variable

$$E(X) = \sum_x \mathrm{p}(x) \cdot x$$

- Example: rolling a dice:

$$E(X) = \sum_{x=1}^{6} \mathrm{p}(x)x = \sum_{x=1}^{6} \frac{x}{6} = 3.5$$

- A function $g(X)$ defines new random variable. In this case:

$$E(g(X)) = \sum_x \mathrm{p}(x)g(x)$$

  Example?
- Also for two random variables:

$$E(X + Y) = E(X) + E(Y)$$

  and if independent

Expectation

- Expectation is a mean (weighted average) of a random variable

$$E(X) = \sum_x \mathrm{p}(x) \cdot x$$

- Example: rolling a dice:

$$E(X) = \sum_{x=1}^{6} \mathrm{p}(x)x = \sum_{x=1}^{6} \frac{x}{6} = 3.5$$

- A function $g(X)$ defines new random variable. In this case:

$$E(g(X)) = \sum_x \mathrm{p}(x)g(x)$$

  Example?

- Also for two random variables:

$$E(X+Y) = E(X) + E(Y)$$

  and if independent

$$E(XY) = E(X)E(Y)$$

Variance

- Variance measures how much values of a random variable vary

$$Var(X) = E[(X - E(X))^2]$$

Variance

- Variance measures how much values of a random variable vary

$$Var(X) = E[(X - E(X))^2]$$

- Standard deviation σ is the square root of the variance

Variance

- Variance measures how much values of a random variable vary

$$Var(X) = E[(X - E(X))^2]$$

- Standard deviation σ is the square root of the variance
- What is the variance of a random variable describing a single throw of a dice?

Entropy

- Entropy is a measure of degree of uncertainty.

Entropy

- Entropy is a measure of degree of uncertainty.
- The most important concept in information theory

Entropy

- Entropy is a measure of degree of uncertainty.
- The most important concept in information theory
- Entropy is a property of a random variable X distributed according the pmf p

Entropy

- Entropy is a measure of degree of uncertainty.
- The most important concept in information theory
- Entropy is a property of a random variable X distributed according the pmf p

$$H(X) = H(\mathrm{p}) = E(-\log_2(\mathrm{p}(x))) = -\sum_x \mathrm{p}(x)\log_2(\mathrm{p}(x))$$

Entropy

- Entropy is a measure of degree of uncertainty.
- The most important concept in information theory
- Entropy is a property of a random variable X distributed according the pmf p

$$H(X) = H(\mathrm{p}) = E(-\log_2(\mathrm{p}(x))) = -\sum_x \mathrm{p}(x) \log_2(\mathrm{p}(x))$$

- For $\log_2(x)$ units are bits, for $\ln(x)$, nats

Entropy as amount of information

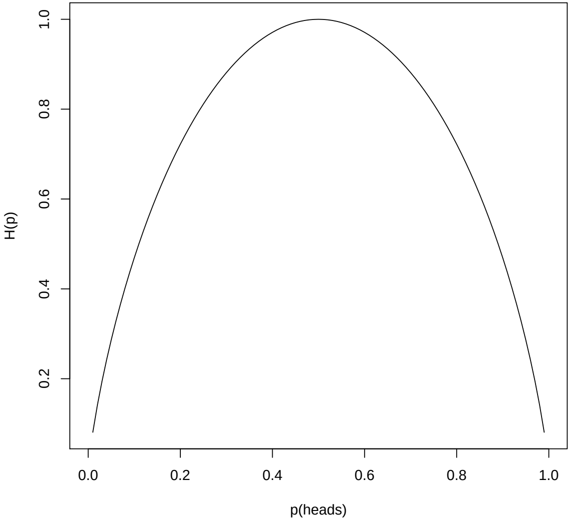
- You can think of entropy as measuring the cost of transmitting information about the result of an experiment
- Fair coin toss:

Entropy as amount of information

- You can think of entropy as measuring the cost of transmitting information about the result of an experiment
- Fair coin toss:

$$
\begin{aligned}
H(X) &= -\sum_{x=0}^{1} \mathrm{p}(x) \log_2(\mathrm{p}(x)) && (1)\\
&= \frac{1}{2}[-\log_2\left(\frac{1}{2}\right) - \log_2\left(\frac{1}{2}\right)] && (2)\\
&= \frac{1}{2} \cdot 2 && (3)
\end{aligned}
$$

Entropy of an unfair coin



Properties of entropy

- $H(p) \geq 0$

Properties of entropy

- $H(p) \geq 0$
- When is entropy $H(p) = 0$?

Properties of entropy

- $H(p) \geq 0$
- When is entropy $H(p) = 0$?
- The highest entropy corresponds to the most uniform distribution

Entropy: joint and conditional

- For two variables X and Y , the amount of information needed to specify values of both

$$H(X, Y) = -\sum_x \sum_y \mathrm{p}(x, y) \log_2(\mathrm{p}(x, y))$$

Entropy: joint and conditional

- For two variables X and Y , the amount of information needed to specify values of both

$$H(X,Y) = -\sum_x \sum_y \mathrm{p}(x,y)\log_2(\mathrm{p}(x,y))$$

- Conditional entropy: if we know the value of X, how much does to cost to transmit the value of Y?

Entropy: joint and conditional

- For two variables X and Y , the amount of information needed to specify values of both

$$H(X,Y) = -\sum_x \sum_y \mathrm{p}(x,y) \log_2(\mathrm{p}(x,y))$$

- Conditional entropy: if we know the value of X, how much does to cost to transmit the value of Y?

$$H(Y|X) = \sum_x \mathrm{p}(x) H(Y|X=x) \tag{4}$$

$$= \sum_x \mathrm{p}(x) \left[-\sum_y \mathrm{p}(y|x) \log(\mathrm{p}(y|x)) \right] \tag{5}$$

$$= -\sum_x \sum_y \mathrm{p}(y|x)\mathrm{p}(x) \log(\mathrm{p}(y|x)) \tag{6}$$

$$= -\sum_{x,y} \mathrm{p}(x,y) \log(\mathrm{p}(y|x)) \tag{7}$$

Conditional entropy: example

- Experiment: a toss of two fair coins
- X: how many heads?
- Y: is there at least one heads?

	X	Y
HH	2	1
HT	1	1
TT	0	0
TH	1	1

What is $H(X)$? What is $H(X|Y)$?

Chain rule for entropy

$$
\begin{aligned}
H(X, Y) &= H(X|Y) + H(Y) \\
H(X_1, \ldots, X_n) &= H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1, \ldots, X_{n-1})
\end{aligned}
$$

Mutual information

- From the chain rule we have

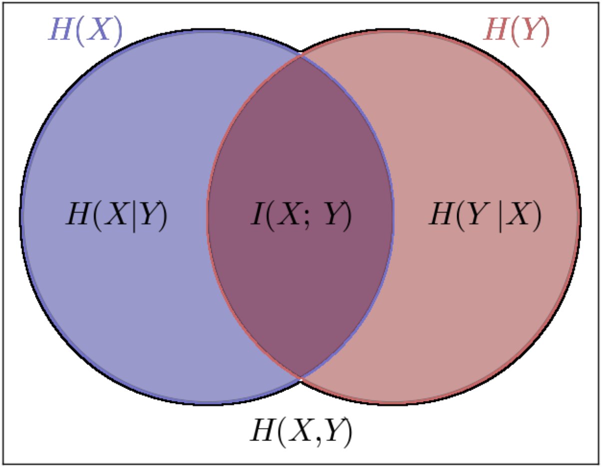
$$H(X) + H(Y|X) = H(Y) + H(X|Y)$$

- Therefore

$$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

- This difference is known as Mutual information $I(X; Y)$
- It measures how much knowing one of the variables reduces uncertainty about the other.

Joint and conditional entropy and mutual information



Mutual information

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= H(X) + H(Y) + H(X,Y) \\
&\dots \\
&= \sum_x \sum_y \mathrm{p}(x,y) \log \left(\frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\mathrm{p}(y)} \right)
\end{aligned}
$$

- What is $H(X|X)$?
- What is $I(X;X)$?

Kullback Leibler divergence

- A measure of the difference between two probability mass functions p and q is Kullback Leibler divergence (relative entropy)

$$D(p||q) = \sum_x \mathrm{p}(x) \log \left(\frac{\mathrm{p}(x)}{\mathrm{q}(x)} \right)$$

- Can be interpreted as an average number of bits wasted by encoding events distributed according to p with a code based on q
- We can define mutual information in terms of KL divergence:

$$I(X;Y) = D(\mathrm{p}(x,y)||\mathrm{p}(x)\mathrm{p}(y))$$

Outline

Bernoulli distribution

- The most basic discrete probability distribution
- Describes the outcome of a single Bernoulli trial
- A Bernoulli trial is an experiment whose outcome is random and can be either of two possible outcomes, success and failure

Bernoulli distribution

- The most basic discrete probability distribution
- Describes the outcome of a single Bernoulli trial
- A Bernoulli trial is an experiment whose outcome is random and can be either of two possible outcomes, success and failure
- If the probability of success is p, then the probability of failure is $1 - p$

Bernoulli distribution

- The most basic discrete probability distribution
- Describes the outcome of a single Bernoulli trial
- A Bernoulli trial is an experiment whose outcome is random and can be either of two possible outcomes, success and failure
- If the probability of success is p, then the probability of failure is $1 - p$
- For example, a single toss of a (possibly biased) coin

The probability mass function of the **Bernoulli** distribution is

$$\mathrm{Bernoulli}(k;p) = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases}$$

The probability mass function of the **Bernoulli** distribution is

$$\mathrm{Bernoulli}(k; p) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases}$$

It can also be expressed as

The probability mass function of the **Bernoulli** distribution is

$$\mathrm{Bernoulli}(k;p) = \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \end{cases}$$

It can also be expressed as

$$\mathrm{Bernoulli}(k;p) = p^k(1-p)^{(1-k)} \text{ for } k \in \{1,0\}$$

The probability mass function of the **Bernoulli** distribution is

$$\mathrm{Bernoulli}(k;p) = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases}$$

It can also be expressed as

$$\mathrm{Bernoulli}(k;p) = p^k(1-p)^{(1-k)} \text{ for } k \in \{1,0\}$$

- What is the expectation of random variable distributed according to Bernoulli?

Binomial distribution

- One of the most important discrete probability distributions

Binomial distribution

- One of the most important discrete probability distributions
- Describes the outcome of a series of Bernoulli trials

Binomial distribution

- One of the most important discrete probability distributions
- Describes the outcome of a series of Bernoulli trials
- For example, a series of tosses of a (possibly biased) coin

Binomial distribution

$$\mathrm{Binomial}(r, n; p) = \binom{n}{r} p^r (1-p)^{n-r}$$

where

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}, \; 0 \leq r \leq n$$

- $\mathrm{Binomial}(r, n; p)$ describes the probability of getting exactly r successes in n trials if the probability of success in an individual trial is p
- $\binom{n}{r}$ is the number of different orders in which we can get r successes in n trials
- Each attempt is independent, so we multiply p r times (successes) and $(1-p)$, $n-r$ times (failures)
- What is the probability of getting **at most** r successes?

Binomial distribution

$$\text{Binomial}(r, n; p) = \binom{n}{r} p^r (1-p)^{n-r}$$

where

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}, \ 0 \leq r \leq n$$

- $\text{Binomial}(r, n; p)$ describes the probability of getting exactly r successes in n trials if the probability of success in an individual trial is p
- $\binom{n}{r}$ is the number of different orders in which we can get r successes in n trials
- Each attempt is independent, so we multiply p r times (successes) and $(1-p)$, $n-r$ times (failures)
- What is the probability of getting **at most** r successes?

$$\sum_{k=0}^{r} \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial test example

- We have made an improvement to our POS tagging model.
- We run the old model and the new model on test sentences.
- The accuracy of the new model is better, but
  - Is it because the system is better? If we repeated the experiment on many other test sentences, would be also get improved accuracy?
  - Or maybe we got an improvement by chance

Null hypothesis

- Use binomial distribution to answer this question
- Focus on the tokens (words) where one of the models makes a mistake and the other gets the right answer
- There are 10 such cases. In 7 cases the new system is better.
- Assume that the new system is actually no better, and that the chance of it being better on any one word is pure chance, 0.5. This is the null hypothesis.
  - How likely are we to get **at least** 7 out of 10 better, given the null hypothesis?
  - How about 60 out of 100? 550 out of 1000?

Null hypothesis

- Use binomial distribution to answer this question
- Focus on the tokens (words) where one of the models makes a mistake and the other gets the right answer
- There are 10 such cases. In 7 cases the new system is better.
- Assume that the new system is actually no better, and that the chance of it being better on any one word is pure chance, 0.5. This is the null hypothesis.
    - How likely are we to get **at least** 7 out of 10 better, given the null hypothesis?
    - How about 60 out of 100? 550 out of 1000?
    - 0.172, 0.028, 0.00086
    - In R: pbinom(7, 10, prob=0.5)

Null hypothesis

- Use binomial distribution to answer this question
- Focus on the tokens (words) where one of the models makes a mistake and the other gets the right answer
- There are 10 such cases. In 7 cases the new system is better.
- Assume that the new system is actually no better, and that the chance of it being better on any one word is pure chance, 0.5. This is the null hypothesis.
  - How likely are we to get **at least** 7 out of 10 better, given the null hypothesis?
  - How about 60 out of 100? 550 out of 1000?
  - 0.172, 0.028, 0.00086
  - In R: pbinom(7, 10, prob=0.5)
- Two-tailed test:
  - Actually we should consider both getting **at least 7 out of 10** or **at most 3 out of 10**