Natural Language Processing

Statistical Inference

Joakim Nivre

Uppsala University
Department of Linguistics and Philology
joakim.nivre@lingfil.uu.se
Statistical Inference

- Inference from a finite set of observations (a sample) to a larger set of unobserved instances (a population or model)
- Two main kinds of statistical inference:
  1. Estimation
  2. Hypothesis testing
- In natural language processing:
  - Estimation – learn model parameters (probability distributions)
  - Hypothesis tests – assess statistical significance of test results
Random Variables

A random variable is a function $X$ that partitions the sample space $\Omega$ by mapping outcomes to a value space $\Omega_X$.

The probability function can be extended to variables:

$$P(X = x) = P(\{\omega \in \Omega \mid X(\omega) = x\})$$

Examples:

1. The part-of-speech of a word $X : \Omega \rightarrow \{\text{noun, verb, adj, \ldots}\}$
2. The number of words in a sentence $Y : \Omega \rightarrow \{1, 2, 3, \ldots\}$.

When we are not interested in particular values, we write $P(X)$. 
Expectation

- The expectation $E[X]$ of a (discrete) numerical variable $X$ is:

$$E[X] = \sum_{x \in \Omega_X} x \cdot P(X = x)$$

- Example: The expectation of the sum $Y$ of two dice:

$$E[Y] = \sum_{y=2}^{12} y \cdot P(Y = y) = \frac{252}{36} = 7$$
Entropy

The entropy $H[X]$ of a discrete random variable $X$ is:

$$H[X] = E[- \log_2 P(X)] = - \sum_{x \in \Omega_X} P(X = x) \log_2 P(X = x)$$

Entropy can be seen as the expected amount of information (in bits), or as the difficulty of predicting the variable.

- Sum of two dice: $- \sum_{y=2}^{12} P(Y = y) \log_2 P(Y = y) \approx 3.27$
- 11-sided die (2–12): $- \sum_{z=2}^{12} \frac{1}{11} \log_2 \frac{1}{11} \approx 3.46$
Quiz 1

Let $X$ be a random variable that maps (English) words to the number of characters they concern

- For example, $X(\text{run}) = 3$ and $X(\text{amok}) = 4$

Which of the following statements do you think are true:

1. $P(X = 0) = 0$
2. $P(X = 5) < P(X = 50)$
3. $E[X] < 50$
Statistical Samples

- A random sample of a variable $X$ is a vector $(X_1, \ldots, X_N)$ of independent variables $X_i$ with the same distribution as $X$
  - It is said to be i.i.d. = independent and identically distributed
  - In practice, it is often hard to guarantee this
  - Observations may not be independent (not i.)
  - Distribution may be biased (not i.d.)

- What is the intended population?
  - A Harry Potter novel is a good sample of J.K. Rowling, or fantasy fiction, but not of scientific prose
  - This is relevant for domain adaptation in NLP
Estimation

- Given a random sample of $X$, we can define sample variables, such as the sample mean:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

- Sample variables can be used to estimate model parameters (population variables)
  1. Point estimation: use variable $X$ to estimate parameter $\phi$
  2. Interval estimation: use variables $X_{\text{min}}$ and $X_{\text{max}}$ to construct an interval such that $P(X_{\text{min}} < \phi < X_{\text{max}}) = p$, where $p$ is the confidence level adopted
Maximum Likelihood Estimation (MLE)

- Likelihood of parameters $\theta$ given sample $x_1, \ldots, x_N$:

$$L(\theta|x_1, \ldots, x_N) = P(x_1, \ldots, x_N|\theta) = \prod_{i=1}^{N} P(x_i|\theta)$$

- Maximum likelihood estimation – choose $\theta$ to maximize $L$:

$$\max_{\theta} L(\theta|x_1, \ldots, x_N)$$

- Basic idea:
  - A good sample should have a high probability of occurring
  - Thus, choose the estimate that maximizes sample probability
Examples

- Sample mean is an MLE of expectation:
  \[ \hat{E}[X] = \bar{X} \]
  - For example, estimate expected sentence length in a certain type of text by mean sentence length in a representative sample

- Relative frequency is an MLE of probability:
  \[ \hat{P}(X = x) = \frac{f(x)}{N} \]
  - For example, estimate the probability of a word being a noun by the relative frequency of nouns in a suitable corpus
MLE for Different Distributions

- Joint distribution of $X$ and $Y$:
  \[
  \hat{P}_{\text{MLE}}(X = x, Y = y) = \frac{f(x, y)}{N}
  \]

- Marginal distribution of $X$:
  \[
  \hat{P}_{\text{MLE}}(X = x) = \sum_{y \in \Omega_Y} \hat{P}_{\text{MLE}}(X = x, Y = y)
  \]

- Conditional distribution of $X$ given $Y$:
  \[
  \hat{P}_{\text{MLE}}(X = x | Y = y) = \frac{\hat{P}_{\text{MLE}}(X = x, Y = y)}{\hat{P}_{\text{MLE}}(Y = y)}
  \]
  \[
  = \frac{\hat{P}_{\text{MLE}}(X = x, Y = y)}{\sum_{x \in \Omega_X} \hat{P}_{\text{MLE}}(X = x, Y = y)}
  \]
Consider the following sample of English words:

\{once, upon, a, time, there, was, a, frog\}

What is the MLE of word length (number of characters) based on this sample?

1. 4
2. 8
3. 3.25