

Statistical Methods

GSLT Course, Week 1

Joakim Nivre

Outline

Introduction

Probability Theory

Stochastic Variables

Statistical Inference

Language Modeling

Outline

Introduction

Probability Theory

Stochastic Variables

Statistical Inference

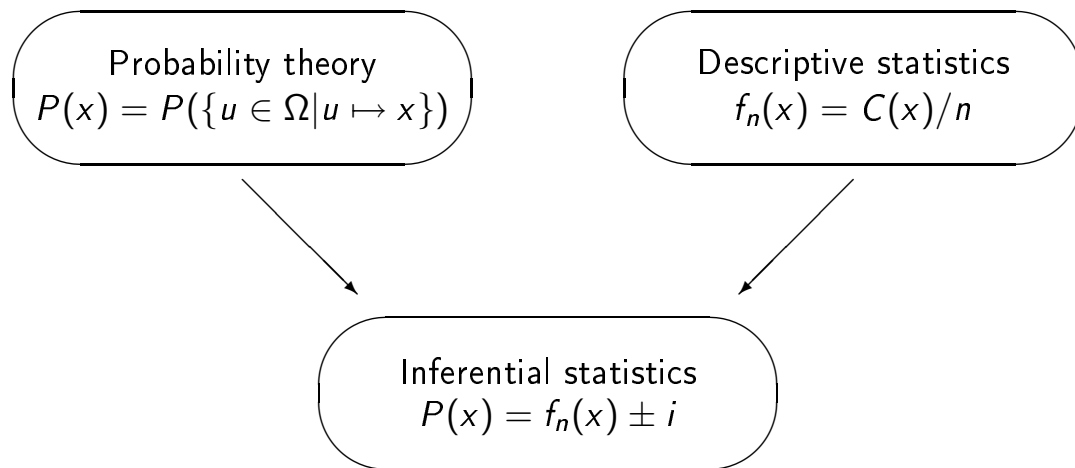
Language Modeling

Introduction

► Basic questions:

1. What is statistics?
2. How can it be used in natural language processing?

Elements of statistical methods



Probability theory

- ▶ Mathematical theory of uncertainty
 - ▶ Random experiments = events where we cannot predict with certainty what is going to happen
 - ▶ Examples: tossing a coin, tomorrow's weather
- ▶ Used for building models of stochastic systems
 - ▶ Stochastic = not deterministic
 - ▶ Stochastic systems modeled as systems of random experiments
- ▶ Human language is a stochastic system?

Descriptive statistics

- ▶ Methods for summarizing (large) data sets
- ▶ Example:
 - ▶ A corpus contains 4321 sentences and 56789 words.
 - ▶ The average sentence length is 13.14 words.
 - ▶ The shortest sentence is 1 word long.
 - ▶ The longest sentence is 98 words long.
- ▶ The most common statistical measures are
 - ▶ frequency counts,
 - ▶ averages,
 - ▶ measures of variation.

Inferential statistics

- ▶ Methods for drawing inferences from (large) datasets
- ▶ Two main inference types:
 1. Estimation: Use data to guess (predict) the value of some unknown quantity (population variable)
 2. Hypothesis testing: Use data to corroborate or refute hypotheses about these quantities
- ▶ Examples:
 - ▶ Predicting the average sentence length in all of Swedish newspaper text from a sample of 1000 articles (estimation)
 - ▶ Testing the hypothesis that the average sentence length in Swedish newspaper text has increased from 1950 to 2000, using one sample from 1950 and one sample from 2000 (hypothesis testing)

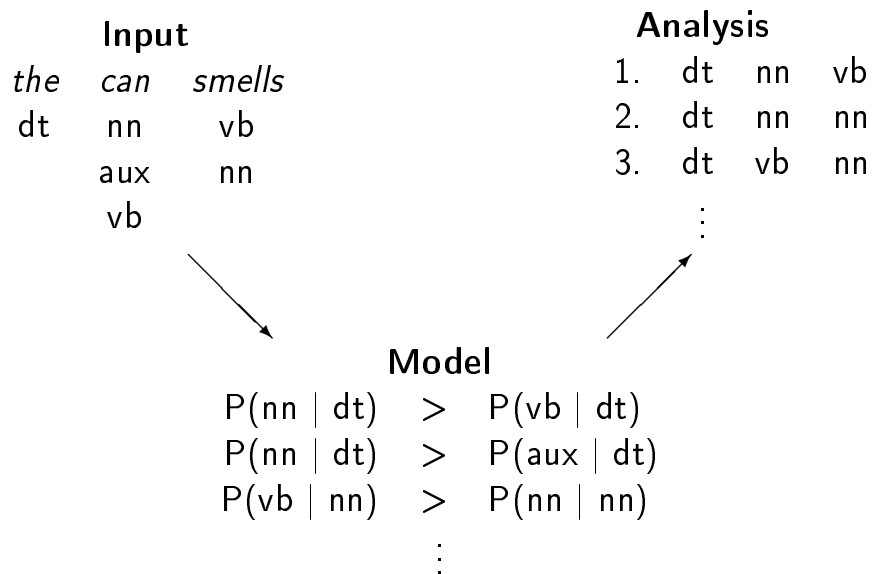
Relevance for Language Technology

- ▶ Three main applications of statistics:
 1. Processing: Use statistical models to process natural language input or output
 2. Learning: Use inferential statistics to learn from examples (corpus data)
 3. Evaluation: Use statistics to assess the performance of language processing systems

Language Processing

- ▶ The most common use of statistics in language processing is to use probability models for disambiguation by ranking alternative analyses with regard to their probability (or related measures such as entropy).
- ▶ Examples:
 1. N-gram language models in speech recognition
 2. Part-of-speech tagging using hidden Markov models
 3. Syntactic parsing using stochastic grammars
 4. Word sense disambiguation using Bayesian classifiers

Part-of-Speech Tagging



Language Learning

- ▶ Statistical inference can be used to construct theoretical models from linguistic data.
- ▶ Examples:
 1. Estimating probabilities for part-of-speech tagging
 2. Extracting translation equivalents from parallel corpora
 3. Extracting collocations from corpora

Part-of-Speech Tagging Again

- ▶ The probability that the word *can* realizes the part-of-speech noun (nn) can be estimated with the relative frequency in a corpus:

$$P(\text{can}|\text{nn}) \approx \frac{C(\text{can}, \text{nn})}{C(\text{nn})}$$

- ▶ The probability that a noun (nn) immediately follows a determiner (dt) can be estimated analogously:

$$P(\text{nn}|\text{dt}) \approx \frac{C(\text{dt}, \text{nn})}{C(\text{dt})}$$

Statistical Learning

- ▶ Supervised learning:
 - ▶ Statistical estimation from labeled data sample (training data)
 - ▶ Example: $P(x) = f_n(x)$ (MLE)
 - ▶ Smoothing to counter the effect of sparse data
- ▶ Unsupervised learning:
 - ▶ Learning from raw data (more difficult)
 - ▶ Often based on numerical optimization
 - ▶ Example:
 1. Guess initial estimates, e.g., $P(x) = \frac{1}{n}$
 2. Expectation-maximization to iteratively improve estimates

Evaluation

- ▶ Statistics may be used in three ways in empirical evaluation of language processing systems:
 1. Descriptive statistics may be used to get quantitative measurements of quality, e.g. error rate, precision-recall, running time.
 2. Estimation may be used to derive confidence intervals for quantitative measurements.
 3. Hypothesis testing may be used to determine whether the measured difference between two different systems is significant or not.

Part-of-Speech Tagging Once More

- ▶ A tagger T_1 is tested on a sample of 1000 words and gets an accuracy rate of 0.92 (92%). How precise is this measurement?

$$0.92 \pm 1.96 \sqrt{\frac{0.92 \cdot 0.08}{1000}} = 0.92 \pm 0.017$$

- ▶ A second tagger T_2 is tested on the same sample and gets an accuracy rate of 0.94. The variance of the difference between T_1 and T_2 is 0.13. Is T_2 significantly better?

$$t = \frac{0.02}{\sqrt{\frac{0.13}{1000}}} = 1.75 < 1.96$$

Summary – Introduction

- ▶ What is statistics?
 1. Probability theory
 2. Descriptive statistics
 3. Inferential statistics
- ▶ How can statistics be used in language technology?
 1. Probability models for language processing
 2. Statistical learning to build processing systems
 3. Statistical evaluation of processing systems

Outline

Introduction

Probability Theory

Stochastic Variables

Statistical Inference

Language Modeling

Probability Theory

- ▶ Topics:
 1. The notion of probability
 2. Probability models
 3. Simple and conditional probability
- ▶ Focus on basic ideas rather than technical details

The Notion of Probability

- ▶ The probability of A , $P(A)$, is a real number between 0 and 1:
 1. If $P(A) = 0$, then A is impossible (never happens).
 2. If $P(A) = 1$, then A is necessary (always happens).
 3. If $0 < P(A) < 1$, then A is possible (may happen).
- ▶ But what does it mean?
 1. Classical theory: #Favorable cases / #Possible cases
 2. Probabilities as relative frequencies
 3. Probabilities as subjective beliefs

Sample Spaces

- ▶ Sample space Ω = the set of all possible elementary outcomes.
 1. Throw a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
 2. Flip a coin: $\Omega = \{\text{head}, \text{tails}\}$
 3. Measure the temperature: $\Omega = \{u | u \in R, u > -273\}$
 4. Utter a word: $\Omega = \{u | u \text{ is a word}\}$
- ▶ Sample spaces may be discrete (1, 2, 4) or continuous (3).

Events

- ▶ Event = subset of the sample space Ω .
 1. Throw a die: At least 4 = $\{4, 5, 6\}$
 2. Flip a coin: Neither head nor tails = \emptyset
 3. Temperature: Above freezing = $\{u | u \in R, u > 0\}$
 4. Uttering a word: Uttering a noun = $\{u | u \text{ is a noun}\}$
- ▶ For discrete sample spaces, the event space is the power set 2^Ω of the sample space.

Composite Experiments

- ▶ In many situations, we are interested in complex phenomena that can be decomposed into simpler ones.
 1. Throwing two dice: (Throwing one die, Throwing one die)
 2. Three-word utterance: (One word, One word, One word)
- ▶ If experiments e_1, \dots, e_n have sample spaces $\Omega_1, \dots, \Omega_n$, then the composite experiment (e_1, \dots, e_n) has the sample space $\Omega_1 \times \dots \times \Omega_n$.
 1. Throwing two dice: $\Omega = \{(m, n) | 1 \leq m, n \leq 6\}$
 2. Three-word utterance: $\Omega = \{(u, v, w) | u, v, w \text{ are words}\}$

Axioms of Probability

- ▶ $P(A)$ = The probability of event A .
- ▶ Axioms:
 1. $P(A) \geq 0$
 2. $P(\Omega) = 1$
 3. If A and B are disjoint, then $P(A \cup B) = P(A) + P(B)$

Simple Probability: Examples

- ▶ The sum N of two dice:
 1. $P(N > 9) = 1/6$
 2. $P(N < 4) = 1/12$
 3. $P(N < 4 \cup N > 9) = 1/4$
- ▶ Part-of-speech of arbitrary word W in text:
 1. $P(W \text{ noun}) = 0.15$
 2. $P(W \text{ verb}) = 0.1$
 3. $P(W \text{ noun} \cup W \text{ verb}) = 0.25$

Some Useful Theorems

1. $P(\Omega - A) = 1 - P(A)$
2. $P(\emptyset) = 0$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4. $P(A \cap B) \leq P(A \cup B) \leq P(A) + P(B)$
5. If $A \subseteq B$, then $P(B - A) = P(B) - P(A)$
6. If A_1, \dots, A_n is a partitioning of Ω , then

$$P(B) = P(A_1 \cap B) + \dots + P(A_n \cap B)$$

Conditional Probability

- ▶ The probability of A given B :

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- ▶ Multiplication rule:

$$P(A \cap B) = P(A)P(B|A)$$

- ▶ Bayes law:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Conditional Probability: Example 1

- ▶ Sum N of two dice:
 1. $P(N \text{ even} \cap N > 9) = 1/9$
 2. $P(N \text{ even}) = 1/2$
 3. $P(N > 9) = 1/6$
 4. $P(N > 9|N \text{ even}) = 2/9$
 5. $P(N \text{ even}|N > 9) = 2/3$

Conditional Probability: Example 2

- ▶ Arbitrary word W from English text:
 - ▶ Assume:
 1. $P(W = \textit{that}) = 0.01$
 2. $P(W \text{ pronoun}) = 0.1$
 3. $P(W = \textit{that} \cap W \text{ pronoun}) = 0.008$
 - ▶ Then we have:
 1. $P(W = \textit{that} | W \text{ pronoun}) = 0.008/0.1 = 0.08$
 2. $P(W \text{ pronoun} | W = \textit{that}) = 0.008/0.01 = 0.8$

Independence

- ▶ If A and B are independent events, then
 1. $P(A \cap B) = P(A)P(B)$
 2. $P(A|B) = P(A)$
 3. $P(B|A) = P(B)$

Independence: Example 1

- ▶ Sum of two dice:
 1. $P(\text{first die} = 1) = 1/6$
 2. $P(N \text{ even}) = 1/2$
 3. $P(\text{first die} = 1 \text{ and } N \text{ even}) = 1/12$
 4. $P(N \text{ even} | \text{first die} = 1) = 1/2$
 5. $P(\text{first die} = 1 | N \text{ even}) = 1/6$
- ▶ The events are independent.

Independence: Example 2

- ▶ Arbitrary word W from English text:
 1. $P(W = \textit{that}) = 0.01$
 2. $P(W \text{ pronoun}) = 0.1$
 3. $P(W = \textit{that} \cap W \text{ pronoun}) = 0.008$
- ▶ The events are not independent.

Summary – Probability Theory

- ▶ Probability model:
 1. Sample space Ω
 2. Event space E
 3. Probability function P
- ▶ Important notions:
 1. Simple and conditional probability
 2. Independence

Outline

Introduction

Probability Theory

Stochastic Variables

Statistical Inference

Language Modeling

Stochastic Variables

- ▶ Topics:
 1. Stochastic variables
 2. Probability functions for variables
 3. Expectation and variance for numerical variables
 4. Basic information theory
- ▶ Still focus on basic ideas rather than technical details

Stochastic Variables

- ▶ A stochastic variable X is a function from a sample space Ω (the domain of X) to a value space Ω_X (the range of X).
- ▶ Examples:
 1. The part-of-speech of an arbitrary word from a corpus is a stochastic variable X with $\Omega = \{w | w \text{ is a corpus token}\}$ and $\Omega_X = \{\text{noun, verb, adjective, ...}\}$.
 2. The sum of two dice is a stochastic variable Y with $\Omega = \{(x, y) | 1 \leq x, y \leq 6\}$ and $\Omega_Y = \{z | 2 \leq z \leq 12\}$.

Different Kinds of Variables

- ▶ If Ω_X is a subset of the set of real numbers, then X is said to be numerical; otherwise it is categorical.
- ▶ If Ω_X is finite or countably infinite, then X is said to be discrete.
- ▶ Examples:
 1. The part-of-speech X of an arbitrary word from a corpus is a discrete, categorical variable, since Ω_X is finite and not numerical.
 2. The sum Y of two dice is a discrete numerical variable, since Ω_Y is finite and numerical.

Frequency Functions

- ▶ The probability $P(X = x)$ of variable X assuming value x is given by the frequency function f_X :

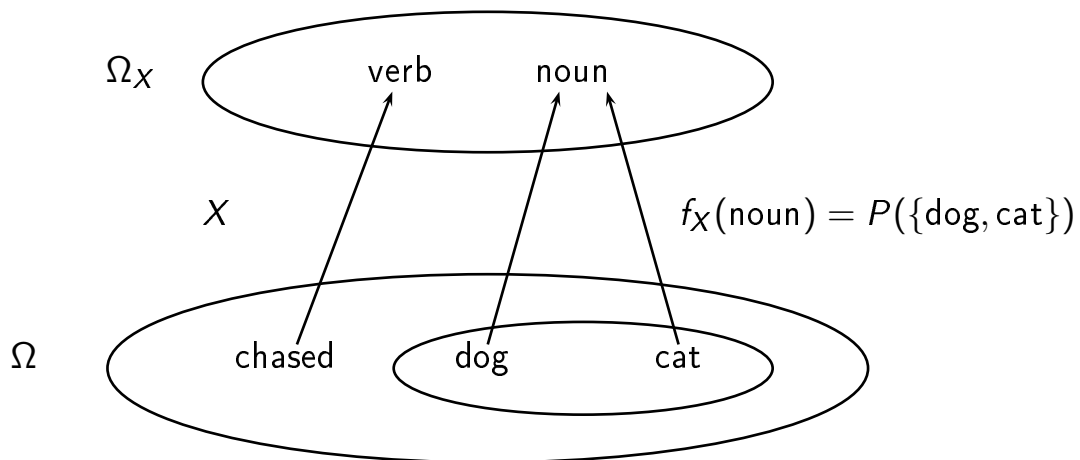
$$f_X(x) = P(X = x)$$

- ▶ For discrete variables, this is equivalent to summing the probability of all outcomes in Ω that are mapped to x by X :

$$f_X(x) = P(\{u \in \Omega \mid X(u) = x\}) = \sum_{u: X(u)=x} P(u)$$

- ▶ Example: The probability of sampling a noun from a corpus can be obtained by summing the probability of sampling each individual noun token in the corpus.

Variables and Abstraction



Distribution Functions

- ▶ For numerical variables there is also a (cumulative) distribution function F_X :

$$F_X(x) = P(X \leq x)$$

- ▶ For discrete numerical variables, the distribution function can be related to the frequency function as follows:

$$F_X(x) = \sum_{y \leq x} f_X(y)$$

- ▶ Example: The probability of getting a sum of at least 4 with two dice can be obtained by summing the probability of the sums 2, 3 and 4.

Expectation

- ▶ Let X be a discrete numerical variable with value space Ω_X . The expectation of X , $E[X]$, is defined as follows:

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

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

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

Variance

- ▶ Let X be a discrete stochastic variable with expectation μ . The variance of X , $\text{Var}[X]$, is defined as follows:

$$\text{Var}[X] = E[(X - \mu)^2] = \sum_{x \in \Omega_X} (x - \mu)^2 \cdot f_X(x)$$

- ▶ Example: The variance of the sum Y of two dice:

$$\text{Var}[Y] = \sum_{y=2}^{12} (y - 7)^2 \cdot f_Y(y) = \frac{210}{36} \approx 5.83$$

Entropy

- ▶ Let X be a discrete stochastic variable. The entropy of X , $H[X]$, is defined as follows:

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

- ▶ Example: The entropy of the sum Y of two dice:

$$H[Y] = - \sum_{y=2}^{12} f_Y(y) \log_2 f_Y(y) \approx 3.27$$

More on Entropy

- ▶ The entropy of a variable X can be interpreted as the expected amount of information (measured in bits) when learning the value of X :

$$I_X(x) = -\log_2 f_X(x)$$

- ▶ Given a finite value space Ω_X of size n , entropy is maximized if $f_X(x) = \frac{1}{n}$ for all $x \in \Omega_X$.
- ▶ Example: Entropy of the outcome Z of an 11-sided die (2–12):

$$H[Z] = - \sum_{z=2}^{12} \frac{1}{11} \log_2 \frac{1}{11} \approx 3.46$$

Joint and Conditional Probability

- Let X and Y be stochastic variables with sample spaces Ω_1 and Ω_2 and value spaces Ω_X and Ω_Y , respectively.

1. The joint probability of X and Y is given by their joint probability function $f_{(X,Y)}$:

$$\begin{aligned} f_{(X,Y)}(x,y) &= P(X = x, Y = y) \\ &= P(\{(u,v) \in \Omega_1 \times \Omega_2 | X(u) = x, Y(v) = y\}) \end{aligned}$$

2. The conditional probability of X given Y is given by the conditional probability function $f_{X|Y}$:

$$f_{X|Y}(x|y) = P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

Joint and Conditional Entropy

- Let X and Y be stochastic variables with value spaces Ω_X and Ω_Y , respectively.

1. Joint entropy of X and Y :

$$H[X, Y] = - \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} f_{(X,Y)}(x,y) \log_2 f_{(X,Y)}(x,y)$$

2. Conditional entropy of X given Y :

$$H[X|Y] = - \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} f_{(X,Y)}(x,y) \log_2 f_{X|Y}(x|y)$$

Stochastic Vectors

- ▶ The notions of joint and conditional probability can be generalized to arbitrary vectors (sequences) of variables:
 1. Joint probability:

$$P(X_1 = x_1, \dots, X_n = x_n) = P(\{(u_1, \dots, u_n) \in \Omega_1 \times \dots \times \Omega_n \mid X_1(u_1) = x_1, \dots, X_n(u_n) = x_n\})$$

2. Conditional probability:

$$\frac{P(X_1 = x_1, \dots, X_n = x_n \mid Y_1 = y_1, \dots, Y_m = y_m)}{P(Y_1 = y_1, \dots, Y_m = y_m)}$$

Independence

- ▶ Stochastic variables X_1, \dots, X_n (defined on the same underlying sample space) are independent if and only if the following holds for all $(x_1, \dots, x_n) \in \Omega_{X_1} \times \dots \times \Omega_{X_n}$:

$$P(X_1 = x_1, \dots, X_n = x_n) = P(X_1 = x_1) \cdot \dots \cdot P(X_n = x_n)$$

- ▶ Corollary: If X and Y are independent variables then the following conditions hold (for all $x \in \Omega_X$ and $y \in \Omega_Y$):
 1. $P(X = x \mid Y = y) = P(X = x)$
 2. $P(Y = y \mid X = x) = P(Y = y)$

Part-of-Speech Bigrams 1

- ▶ Let (X_1, X_2) be the parts-of-speech of an arbitrary bigram and let the following probabilities be given:
 1. $P(X_1 = \text{noun}) = P(X_2 = \text{noun}) = 0.2$
 2. $P(X_1 = \text{adj}) = P(X_2 = \text{adj}) = 0.05$
 3. $P(X_1 = \text{det} | X_2 = \text{noun}) = 0.3$
 4. $P(X_1 = \text{det} | X_2 = \text{adj}) = 0.6$
 5. $P(X_1 = \text{det} | X_2 \notin \{\text{noun}, \text{adj}\}) = 0$
- ▶ Question: What is $P(X_2 = \text{noun} | X_1 = \text{det})$?

Part-of-Speech Bigrams 2

- ▶ Using Bayes' law:

$$\frac{P(X_2 = \text{noun}) \cdot P(X_1 = \text{det} | X_2 = \text{noun})}{P(X_1 = \text{det})}$$

- ▶ Using the law of total probability:

$$\frac{P(X_2 = \text{noun}) \cdot P(X_1 = \text{det} | X_2 = \text{noun})}{P(X_1 = \text{d} | X_2 = \text{n}) \cdot P(X_2 = \text{n}) + P(X_1 = \text{d} | X_2 = \text{a}) \cdot P(X_2 = \text{a})}$$

- ▶ Putting in the numbers:

$$\frac{0.2 \cdot 0.3}{0.3 \cdot 0.2 + 0.6 \cdot 0.05} = 0.67$$

Part-of-Speech Bigrams 3

- ▶ Consider:
 1. $P(X_1 = \text{det}) = 0.09$
 2. $P(X_2 = \text{noun}) = 0.2$
 3. $P(X_1 = \text{d}, X_2 = \text{n}) = P(X_1 = \text{d}) \cdot P(X_2 = \text{n} | X_1 = \text{d}) = 0.06$
 4. $P(X_1 = \text{det}) \cdot P(X_2 = \text{noun}) = 0.2 \cdot 0.09 = 0.018$
 5. $0.018 \neq 0.06$
- ▶ Conclusion: X_1 and X_2 are not independent variables.

Summary – Stochastic Variables

- ▶ Stochastic variables:
 1. Numeric and categorical
 2. Discrete and continuous
- ▶ Important notions:
 1. Frequency and distribution functions
 2. Expectation, variance and entropy
 3. Joint and conditional probability
 4. Independence

Outline

Introduction

Probability Theory

Stochastic Variables

Statistical Inference

Language Modeling

Statistical Inference

- ▶ Topics:
 1. Sampling
 2. Estimation
 3. Hypothesis testing
- ▶ Basic elements of statistical learning

Sampling

- ▶ Let X be a stochastic variable.
 1. A vector (X_1, \dots, X_n) of independent variables X_i with the same distribution as X is said to be a random sample of X .
 2. A value vector (x_1, \dots, x_n) such that $X_1 = x_1, \dots, X_n = x_n$ in a particular experiment is called a statistical material.
- ▶ Example:
 - ▶ Consider a corpus C consisting of words (w_1, \dots, w_n) .
 - ▶ Can we regard C as a statistical material resulting from a sample (W_1, \dots, W_n) of the word variable W ?
 - ▶ Why (not)?

Sample Variables

- ▶ Given a random sample of a variable X , we can define new stochastic variables that are functions of the sample, called sample variables:
 1. The sample mean: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$
 2. The sample variance: $s_n^2 = \frac{1}{(n-1)} \sum_{i=1}^n (X_i - \bar{X}_n)^2$
- ▶ These variables are called sample variables to distinguish them from the expectation and (true) variance of X , which are called population variables or model parameters.

Statistical Inference

- ▶ Statistical inference is the science of making predictions or inferences from finite sets of observations (samples) to (potentially infinite) sets of new observations (populations).
- ▶ Two main kinds of statistical inference:
 1. Estimation: Use samples and sample variables to predict population variables.
 2. Hypothesis testing: Use samples and sample variables to test hypotheses about populations and population variables.

Estimation

- ▶ Two kinds of estimation:
 1. Point estimation: Use sample variable $f(X_1, \dots, X_n)$ to estimate parameter ϕ .
 2. Interval estimation: Use sample variables $f_1(X_1, \dots, X_n)$ and $f_2(X_1, \dots, X_n)$ to construct an interval such that $P(f_1(X_1, \dots, X_n) < \phi < f_2(X_1, \dots, X_n)) = p$, where p is the confidence level adopted.

Maximum Likelihood Estimation (MLE)

- ▶ Given a statistical material x_1, \dots, x_n and a set of parameters θ , the likelihood function L is:

$$L(x_1, \dots, x_n, \theta) = \prod_{i=1}^n P_{\theta}(x_i)$$

where $P_{\theta}(x_i)$ is the probability that the variable X_i assumes the value x_i given a set of values for the parameters in θ .

- ▶ Maximum likelihood estimation means choosing θ so that the likelihood function is maximized:

$$\max_{\theta} L(x_1, \dots, x_n, \theta)$$

MLE: Example 1

- ▶ Given a random sample (X_1, \dots, X_n) of a numerical variable X , the sample mean \bar{X}_n is a maximum likelihood estimate of the expectation $E[X]$.
- ▶ The average sentence length X in a certain type of text can be estimated with the mean sentence length in a representative sample:

$$\hat{E}[X] = \bar{X}_n$$

MLE: Example 2

- ▶ Given a random sample (X_1, \dots, X_n) of a categorical variable X , the relative frequency of the value x , $f_n(x)$, is a maximum likelihood estimate of the probability $P(X = x)$.
- ▶ The probability of an arbitrary word from a text being a noun can be estimated with the relative frequency of nouns in a suitable corpus of texts:

$$\hat{P}(\text{noun}) = f_n(\text{noun})$$

The Rationale of MLE

- ▶ We want to choose the most probable model given the data:

$$P(\theta|x_1, \dots, x_n) = \frac{P(x_1, \dots, x_n|\theta)P(\theta)}{P(x_1, \dots, x_n)}$$

$$\arg \max_{\theta} P(\theta|x_1, \dots, x_n) = \arg \max_{\theta} P(x_1, \dots, x_n|\theta)P(\theta)$$

- ▶ If we assume a uniform distribution for $P(\theta)$, then

$$\arg \max_{\theta} P(\theta|x_1, \dots, x_n) = \arg \max_{\theta} P(x_1, \dots, x_n|\theta)$$

- ▶ The status of $P(\theta)$ is controversial in statistical theory (Bayesians vs. Frequentists)

Motivating Example (MLE)

- ▶ Assume two sets of two dice:
 1. Two ordinary dice
 2. One die with 5 on all sides; one with 2 on 2 sides, 6 on 4 sides
- ▶ Sums from 5 throws: $\langle 7, 11, 11, 11, 7 \rangle$ (data)
- ▶ Which set of dice was used (model)?
- ▶ Likelihood values:
 1. $P(\text{data}|\text{set 1}) = 1/209952 \approx 0.000005$
 2. $P(\text{data}|\text{set 2}) = 8/243 \approx 0.033$

Practical Considerations (MLE)

- ▶ MLE is a good solution to the estimation problem if the statistical material is large enough. In practice, MLE is often suboptimal because of sparse data.
- ▶ Another problem is caused by the fact that the data sets used in practical applications, such as language corpora, seldom satisfy the conditions for being a true random sample.
- ▶ Practical solutions to the estimation problem often use the MLE as a starting point, applying more or less sophisticated smoothing techniques in an attempt to improve the quality of the estimate.

Interval Estimation

- ▶ In general, we can derive a 95% confidence interval for our maximum likelihood estimate $\hat{\phi}$ of a mean as follows:

$$\hat{\phi} \pm 1.96 \frac{s}{\sqrt{n}}$$

where s is the standard deviation of the underlying variable and n is the number of observations in our statistical material.

- ▶ Examples:

1. Sentence length: $\hat{E}[X] = \bar{X}_n \pm 1.96 \frac{s}{\sqrt{n}}$
2. Noun probability: $\hat{P}(\textit{noun}) = f_n(\textit{noun}) \pm 1.96 \frac{s}{\sqrt{n}}$

More on Interval Estimation

- ▶ The previous formulas for 95% confidence intervals are based on the assumption that the estimated parameter has a normal distribution, an assumption that is theoretically provable in the limit but which is met to varying degrees in the samples used in practice.
- ▶ The formulas also presuppose that the true standard deviance (variance) of the variable is known. In practice, this parameter has to be estimated from the sample as well.

Hypothesis Testing

1. Choose a test statistic t whose distribution is known when the null hypothesis is true.
2. Use t to calculate the probability p of observing the data given that the null hypothesis is true.
3. If $p < \alpha$, reject the null hypothesis, where α is the significance level adopted.

Example: Hypothesis Testing

- ▶ In a random sample of 1000 words from written texts, there were 154 nouns. In a corresponding sample from transcribed spoken dialogues, there were only 128 nouns. Are nouns more frequent in written language?
- ▶ If there is no difference, we should expect $(154 + 128) / 2 = 141$ nouns in a 1000 word sample (the null hypothesis).
- ▶ χ^2 test:
 1. $\chi^2 = \sum \frac{(O-E)^2}{E} = 2.79$
 2. $P(\chi^2 = 2.79)$ with 1 df ≈ 0.1

Summary – Statistical Inference

- ▶ Samples and sample variables
- ▶ Estimation
 1. Point estimation (MLE)
 2. Interval estimation
- ▶ Hypothesis testing: Probability of data given null hypothesis

Outline

Introduction

Probability Theory

Stochastic Variables

Statistical Inference

Language Modeling

Language Modeling

- ▶ A stochastic (or probabilistic) language model M is a model that assigns probabilities to strings in a language L :
 1. $0 \leq P_M(x) \leq 1$ (for all $x \in L$)
 2. $\sum_{x \in L} P_M(x) = 1$
- ▶ Stochastic language models are used in many NLP applications:
 1. Speech recognition
 2. Machine translation
 3. Optical character recognition
 4. Spell checking

Example: Speech Recognition

- ▶ In a typical stochastic speech recognition system, we seek the most probable string of words W for a given acoustic signal S :

$$\arg \max_W P(W|S) = P(W)P(S|W)$$

- ▶ For example, assume that the following three word sequences are equally probable given only the acoustic model ($P(S|W)$):
 1. drank two beers
 2. drank too beers
 3. drank two deers
- ▶ Which interpretation would you prefer? Why?

N-Gram Models

- ▶ In an n -gram model, each word is assumed to be dependent only on $n - 1$ adjacent words:

1. Bigram model ($n = 2$):

$$P(w_1 \cdots w_m) = \prod_{i=1}^m P(w_i | w_{i-1})$$

2. Trigram model ($n = 3$):

$$P(w_1 \cdots w_m) = \prod_{i=1}^m P(w_i | w_{i-2} w_{i-1})$$

Independence Assumptions

- ▶ According to probability theory, the correct model should be:

$$P(w_1 \cdots w_m) = \prod_{i=1}^m P(w_i | w_1 \cdots w_{i-1})$$

- ▶ In an n -gram model, we make the following independence assumption:

$$P(w_i | w_1 \cdots w_{i-1}) = P(w_i | w_{i-(n-1)} \cdots w_{i-1})$$

- ▶ This is equivalent to:

$$P(w_i) = P(w_i | w_j) \text{ for all } j < i - (n - 1)$$

Estimation

- ▶ Given a corpus sampled from the language to be modeled, n -gram probabilities can be estimated as follows:

$$\hat{P}(w_i | w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})}$$

$$\hat{P}(w_i | w_{i-2}w_{i-1}) = \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})}$$

where $C(x)$ is the number of times the string x occurs in the corpus.

- ▶ This is a special case of maximum likelihood estimation (MLE).

Sparse Data and Smoothing

- ▶ Pure MLE is suboptimal for several reasons:
 1. All unseen n -grams will be estimated to have probability zero. This is especially problematic because zero is the annihilator for multiplication.
 2. Estimates for low frequency n -grams are generally unreliable (cf. n -grams with 1, 2, 3 occurrences).
- ▶ Different ways of avoiding zero probabilities and making the estimates for rare n -grams more reliable are known as smoothing (or discounting) methods.

Cardinality Considerations

- ▶ Consider a vocabulary of 100 000 (10^5) words:
 1. The number of distinct bigrams is 10^{10} .
 2. The number of distinct trigrams is 10^{15} .
- ▶ The largest available corpora contain on the order of 10^{10} word tokens, half of which are usually hapax legomena, i.e. words occurring only once.

Smoothing: The Simplest Hack

- ▶ Use MLE but avoid zero probabilities:

$$\hat{P}(w_i | w_{i-2}w_{i-1}) = \begin{cases} \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})} & \text{if } C(w_{i-2}w_{i-1}w_i) > 0 \\ \epsilon & \text{otherwise} \end{cases}$$

- ▶ Drawbacks:
 1. The resulting model is not a correct probabilistic model.
 2. Estimates are still unreliable for low-frequency n -grams.

Additive Smoothing

- ▶ Add k observations to each n -gram count:

$$\hat{P}(w_i | w_{i-2}w_{i-1}) = \frac{C(w_{i-2}w_{i-1}w_i) + k}{C(w_{i-2}w_{i-1}) + k \cdot N_w}$$

where N_w is the number of distinct words in the vocabulary.

- ▶ Common values for k :
 1. Laplace's Law: $k = 1$
 2. Lidstone's Law: $k = 0.5$
- ▶ Resulting model is theoretically correct but tends to overestimate the probability of unseen n -grams.

Held-Out Estimation

- ▶ Divide training data in two parts:
 1. Use first part to divide n -grams into frequency classes, where S_m is the set of n -grams that occur m times.
 2. Use second part to smooth estimates by assigning to each frequency class the probability mass corresponding to its relative frequency in the second subpart of the corpus:

$$\hat{P}(S_m) = \frac{1}{N} \sum_{x \in S_m} C_2(x)$$

- ▶ Held-out estimation can yield very good estimates but reduces the size of the effective training corpus.

Good-Turing Estimation

- ▶ Use expected frequencies of frequencies to reestimate n -gram counts:

$$C^*(x) = (C(x) + 1) \frac{E[|S_{C(x)+1}|]}{E[|S_{C(x)}|]}$$

where $C^*(x)$ is the reestimated count of n -gram x and $E[|S_m|]$ is the expected number of n -grams occurring m times in the training corpus.

- ▶ Problem: How estimate expected frequencies?
- ▶ Good-Turing estimation performs very well for many probabilistic models.

Backoff Smoothing

- ▶ Use MLE for frequent events, back off to a simpler model for rare events:

$$\hat{P}(w_i | w_{i-2}w_{i-1}) = \begin{cases} (1 - d) \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})} & \text{if } C(w_{i-2}w_{i-1}w_i) > t \\ \alpha \cdot \hat{P}(w_i | w_{i-1}) & \text{otherwise} \end{cases}$$

where d is a discounting factor (which can be differentiated for different frequencies using Good-Turing estimation) and α is a renormalization factor.

- ▶ Backoff smoothing (in combination with Good-Turing) is one of the best smoothing methods for n -gram models.

Linear Interpolation

- ▶ Use a weighted sum of progressively simpler models:

$$\hat{P}(w_i | w_{i-2}w_{i-1}) = \lambda_3 \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})} + \lambda_2 \frac{C(w_{i-1}w_i)}{C(w_{i-1})} + \lambda_1 \frac{C(w_i)}{N}$$

where λ_i are coefficients, which may be differentiated for different histories, but which must always sum to 1. Why?

- ▶ Linear interpolation can be a very good smoothing method for n -gram models.

Alternative Models

- ▶ Attempts to develop more advanced probabilistic language models:
 1. Larger n -grams
 2. Class-based models
 3. Grammar-based models
- ▶ The simple bigram and trigram models have proven surprisingly hard to beat.

Practical Assignment

- ▶ N-gram models for artificial blocks world English:
put the red cone on the square
put the block on the blue square on the circle
- ▶ Compare unigram, bigram and trigram models
- ▶ Compare probability estimates to true source probabilities