Review: Probability

BM1: Advanced Natural Language Processing

University of Potsdam

Tatjana Scheffler

tatjana.scheffler@uni-potsdam.de

October 21, 2016

Today

- probability
- random variables
- Bayes' rule
- expectation
- maximum likelihood estimation

Motivations

- Statistical NLP aims to do statistical inference for the field of NL
- *Statistical inference* consists of taking some data (generated in accordance with some unknown *probability distribution*) and then making some inference about this distribution.
- Example: *language modeling* (i.e. how to predict the next word given the previous words)
- Probability theory helps us finding such model

Probability Theory

- How likely it is that something will happen
- Sample space Ω is listing of all possible outcome of an experiment
- Event A is a subset of Ω
- Event space is the powerset of Ω : 2^{Ω}
- Probability function (or distribution):

$$P: 2^{\Omega} \mapsto [0,1]$$

Examples

- An *random variable* X, Y, ... describes the possible outcomes of a random event and the probability of that outcome.
- flip of a fair coin
  - sample space: $\Omega = \{ H , T \}$
  - probabilities of basic outcomes?

a	P(X=a)
H	0.5
T	0.5

- dice roll
  - sample space?
  - probabilities?
- probability distribution of X is the function $a \mapsto P(X{=}a)$

Events

- subsets of the sample space
- atomic events = basic outcomes
- We can assign probability to complex events:
  - $P(X = 1 \text{ or } X = 2)$: prob that X takes value 1 or 2.
  - $P(X \geq 4)$: prob that X takes value 4, 5, or 6.
  - $P(X = 1 \text{ and } Y = 2)$: prob that rv X takes value 1 and rv Y takes value 2.
- In case of language, the sample space is usually finite, i.e. we have *discrete* random variables. There are also continuous rvs.
  - example?

Probability Axioms

- The following axioms hold of probabilities:
  - $0 \leq P(X = a) \leq 1$ for all events $X = a$
  - $P(X \in \Omega) = 1$
  - $P(X \in \varnothing) = 0$
  - $P(X \in A) = P(X = a_1) + \ldots + P(X = a_n)$
    for $A = \{a_1, \ldots, a_n\} \subseteq \Omega$
- Example: If the probability distribution of X is *uniform* with N outcomes,
  i.e. $P(X = a_i) = 1/N$ for all i, then $P(X \in A) = |A| / N$.

Law of large numbers

- Where do we get probabilities from?
  - reasonable assumptions + axioms
  - subjective estimation/postulation
  - law of large numbers
- Law of large numbers: In an infinite number of trials, relative frequency of events converges towards their probabilities

Consequences of Axioms

- The following rules for calculating with probs follow directly from the axioms.
  - Union:

    $P(X \in B \cup C) = P(X \in B) + P(X \in C) - P(X \in B \cap C)$
  - In particular, if B and C are disjoint (and only then), $P(X \in B \cup C) = P(X \in B) + P(X \in C)$
  - Complement:
    $P(X \notin B) = P(X \in \Omega - B) = 1 - P(X \in B)$.
- For simplicity, will now restrict presentation to events $X = a$. Basically everything generalizes to events $X \in B$.

Joint probabilities

- We are very often interested in the probability of two events X = a and Y = b occurring together, i.e. the *joint probability* P(X = a, Y = b).
  - e.g. X = roll of first die, Y = roll of second die
- If we know joint pd, we can recover individual pds by *marginalization*. Very important!

$$P(X = a) = \sum_b P(X = a, Y = b)$$

Conditional Probability

- *Prior probability*: the probability before we consider any additional knowledge: P(X = a)
- Joint probs are trickier than they seem because the outcome of X may influence the outcome of Y.
  - X: draw first card from a deck of 52 cards
    Y: after this, draw second card from deck of cards
  - P(Y is an ace | X is not an ace) = 4/51
    P(Y is an ace | X is an ace) = 3/51
- We write P(Y = a | X = b) for the *conditional probability* that Y has outcome a if we know that X has outcome b.

Conditional and Joint Probability

- $P(X = a, Y = b) = P(Y = b \mid X = a)\, P(X = a)$
  $= P(X = a \mid Y = b)\, P(Y = b)$ (chain rule)
- Thus:

$$P(Y = b \mid X = a) = \frac{P(X = a, Y = b)}{P(X = a)}$$

$$= \frac{P(X = a, Y = b)}{\sum_{b \in B} P(X = a, Y = b)}$$

(marginalization)

(Conditional) independence

- Two events X=a and Y=b are *independent* of each other if :
  - $P(X = a \mid Y = b) = P(X = a)$
  - equivalently: $P(X = a, Y = b) = P(X = a)\ P(Y = b)$
- This means that the outcome of Y has no influence on the outcome of X. Events are *statistically independent*.
  - Typical examples: coins, dice.
- Many events in natural language *not* independent, but we pretend they are to simplify models.

Chain rule, independence

- Chain rule for complex joint events:
  $P(X_1 = a_1, X_2 = a_2, \ldots X_n = a_n)$
  $= P(X_1 = a_1)P(X_2 = a_2 | X_1 = a_1)\ldots P(X_n = a_n | a_1\ldots a_{n-1})$
- In practice, it is typically hard to estimate things like $P(a_n \mid a_1, \ldots, a_{n-1})$ well because not many training examples satisfy complex condition.
- Thus pretend all are independent. Then we have $P(a_1, \ldots, a_n) \approx P(a_1) \ldots P(a_n)$.

Bayes' Theorem

- Important consequence of joint/conditional probability connection
- Bayes' Theorem lets us swap the order of dependence between events
- We saw that $P(Y=b \mid X=a) = \dfrac{P(X=a, Y=b)}{P(X=a)}$
- Bayes' Theorem:

$$P(X=a \mid Y=b) = \frac{P(Y=b \mid X=a) \cdot P(X=a)}{P(Y=b)}$$

Example of Bayes' Rule

- S:stiff neck, M: meningitis
- P(S | M) =0.5, P(M) = 1/50,000 P(S)=1/20
- I have stiff neck, should I worry?

$$P(M \mid S) = \frac{P(S \mid M) P(M)}{P(S)}$$

$$= \frac{0.5 \times 1/50{,}000}{1/20} = 0.0002$$

Expected values / Expectation

- Frequentist interpretation of probability: if P(X = a) = p, and we repeat the experiment N times, then we see outcome "a" roughly p N times.
- Now imagine each outcome "a" comes with reward R(a). After N rounds of playing the game, what reward can we (roughly) expect?
- Measured by *expected value:*

$$E_P[R] = \sum_{a \in A} P(X = a) \cdot R(a)$$

Back to the Language Model

- In general, for language events, P is unknown
- We need to *estimate* P, (or model M of the language)
- We'll do this by looking at evidence about what P must be based on a sample of data (*observations*)

Example: model estimation

- Example: we flip a coin 100 times and observe H 61 times. Should we believe that it is a fair coin?
  - observation: 61x H, 39x T
  - model: assume rv X follows a *Bernoulli* distribution, i.e. X has two outcomes, and there is a value p such that $P(X = H) = p$ and $P(X = T) = 1 - p$.
  - want to estimate the *parameter* p of this model

Estimation of P

- Frequentist statistics
  - parametric methods
  - non-parametric (distribution-free)
- Bayesian statistics

Frequentist Statistics

- Relative frequency: proportion of times an outcome *u* occurs

  $f_u = C(u) / N$

- C(u) is the number of times u occurs in N trials
- For N approaching infinity, the relative frequency tends to stabilize around some number: probability estimates

Non-Parametric Methods

- No assumption about the underlying distribution of the data
- For ex, simply estimate P empirically by counting a large number of random events is a distribution-free method
- Less prior information, more training data needed

Parametric Methods

- Assume that some phenomenon in language is acceptably modeled by one of the well-known family of distributions (such binomial, normal)
- We have an explicit probabilistic model of the process by which the data was generated, and determining a particular probability distribution within the family requires only the specification of a few parameters (less training data)

Binomial Distribution

- Series of trials with only two outcomes, each trial being independent from all the others
- Number *r* of successes out of *n* trials given that the probability of success in any trial is *p*:

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

Normal (Gaussian) Distribution

- Continuous
- Two parameters: mean μ and standard deviation σ

$$n(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Used in clustering

Maximum Likelihood Estimation

- We want to estimate the parameters of our model from frequency observations. There are many ways to do this. For now, we focus on *maximum likelihood estimation*, MLE.
- *Likelihood* L(O ; p) is the probability of our model generating the observations O, given parameter values p.
- Goal: Find value for parameters that maximizes the likelihood.

ML Estimation

- For Bernoulli and multinomial models, it is extremely easy to estimate the parameters that maximize the likelihood:
  - P(X = a) = f(a)
  - in the coin example above, just take p = f(H)
- Why is this?

Bernoulli model

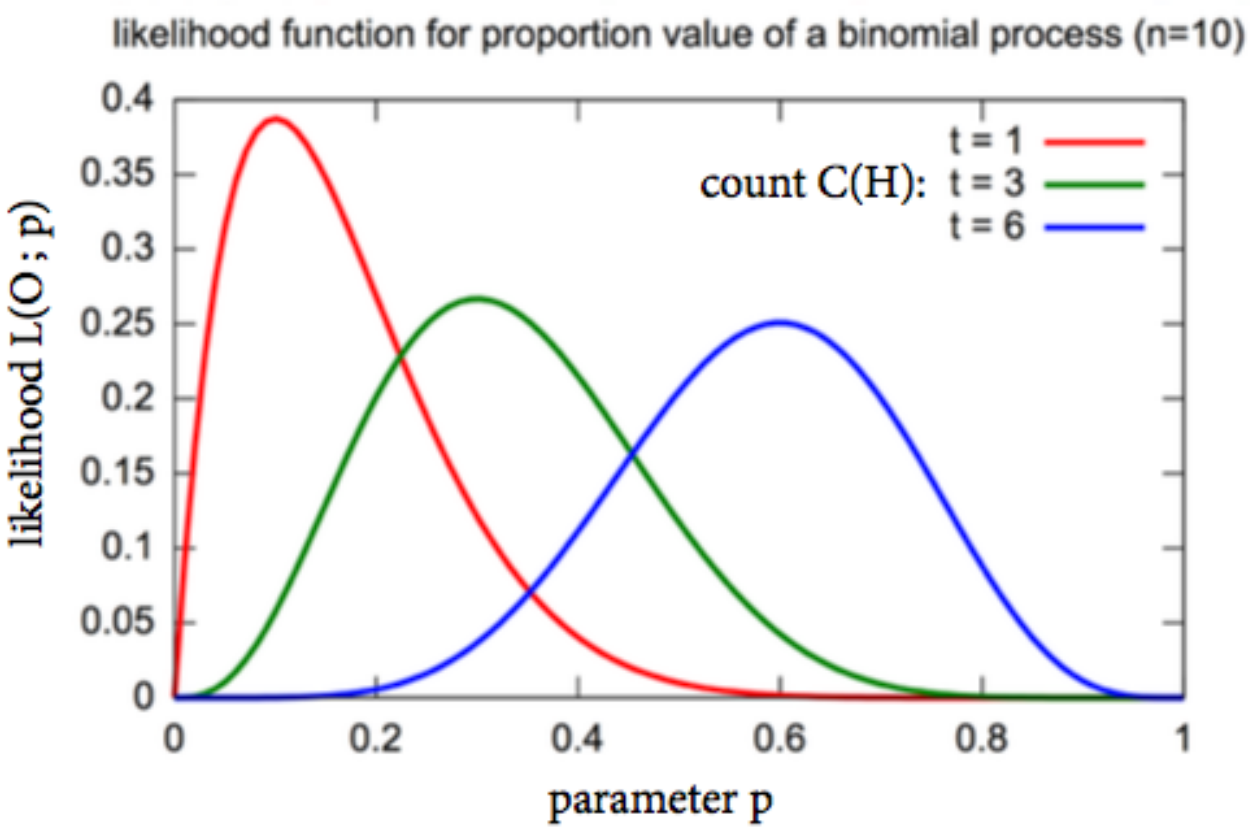
- Let's say we had training data C of size N, and we had N_H observations of H and N_T observations of T.

likelihood $L(C) = \prod_{i=1}^{N} P(w_i \mid p) = \prod_{i=1}^{N} p^{N_H}(1-p)^{N_T}$

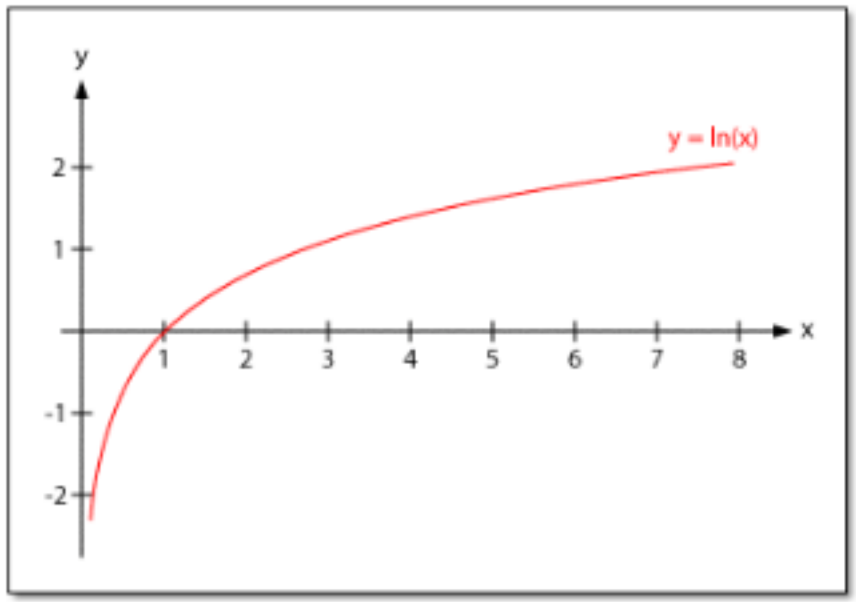
log-likelihood

$$\ell(C) = \log L(C) = \sum_{i=1}^{N} \log P(w_i \mid p) = N_H \log p + N_T \log(1-p)$$

Likelihood functions



Logarithm is monotonic



- Observation: If $x_1 > x_2$, then $\ln(x_1) > \ln(x_2)$.
- Therefore, $\operatorname{argmax}_p L(C) = \operatorname{argmax}_p l(C)$

Maximizing the log-likelihood

- Find maximum of function by setting derivative to zero:

$$\ell(C) = N_H \log p + N_T \log(1-p)$$

$$\frac{d\ell(C)}{dp} = \frac{N_H}{p} - \frac{N_T}{1-p}$$

- Unique solution is $p = N_H / N = f(H)$.

More complex models

- Many, many models we use in NLP are *multinomial* probability distributions. More than two outcomes possible; think dice rolling.
- MLE result generalizes to multinomial models: $P(X = a) = f(a)$.
- Maximizing log-likelihood uses technique called *Lagrange multipliers* to ensure parameters sum to 1.
- If you want to see the details, see Murphy paper on the website.

Conclusion

- Probability theory is essential tool in modern NLP.
- Important concepts today:
  - random variable, probability distribution
  - joint and conditional probs; Bayes' rule; independence
  - expected values
  - statistical models; parameters; likelihood; MLE
- We will use all of these concepts again and again in this course. If you have questions, ask me early.

next Friday

- n-gram models

- (Tuesday: practical session on Python, NLTK, getting ready for assignment 1, etc.)