# Statistical Natural Language Processing

## Statistical models: learning, inference, estimation, prediction

Çağrı Çöltekin

University of Tübingen
Seminar für Sprachwissenschaft

Summer Semester 2017

# Overview

- Many methods/tools we use in NLP can broadly be classified as *statistical models*
- Statistical models have a central role in ML and statistical data analysis
- We will go through an overview of statistical modeling in this lecture

# Models in science and practice

Modeling is a basic activity in science and practice.
A few examples:

- Galilean model of solar system
- Bohr model of atom
- Animal models in medicine
- Scale models of buildings, bridges, cars, …
- Econometric models
- Models of atmosphere

# What do we do with models?

- Inference: learn more about the reality being modeled
  - verify or compare hypotheses on the model
- Prediction: predict the (feature) events/behavior using the model

# Models are not reality

All models are wrong, some are useful.

- All models make some (simplifying) assumptions that do not match with reality
- (some) models are useful despite (or, sometimes, because of) these assumptions / simplifications

Box and Draper (1986, p. 424)

# Statistical models

- Statistical models are mathematical models that take uncertainty into account
- Statistical models are models of data
- We express a statistical model in the form,

$$\text{outcome} = \text{model prediction} + \text{error}$$

- ‘error’ or uncertainty is part of the model description

# Parametric models

Most statistical models are described by a set of parameters $\boldsymbol{w}$

$$y = f(x; \boldsymbol{w}) + \epsilon$$

- $x$ is the input to the model
- $y$ is the quantity or label assigned to for a given input
- $w$ is the parameter(s) of the model
- $f(x; w)$ is the model's estimate ($\hat{y}$) of $y$ given the input $x$
- $\epsilon$ represents the uncertainty or noise that we cannot explain or account for (may include additional parameters)

# Parametric models

$$y = f(x; \boldsymbol{w}) + \epsilon$$

- In machine learning (and in this course), focus is on prediction: given $x$, make accurate predictions of $y$
- In statistics, the focus is on inference (testing hypotheses or explaining the observed phenomena)
  - for example, does $x$ have an effect on $y$?
- For both purposes, finding a good estimate $\boldsymbol{w}$ is important
- For inference, properties of $\epsilon$ (e.g., its distribution and variance) is important

# What are good estimates / estimators?

*Bias* of an estimate is the difference between the value being estimated, and the expected value of the estimate

$$\mathsf{B}(\hat{w}) = \mathsf{E}[\hat{w}] - w$$

- An *unbiased* estimator has 0 bias

*Variance* of an estimate is, simply its variance, the value of the squared deviations from the mean estimate

$$var(\hat{w}) = \mathsf{E}\left[(\hat{w} - \mathsf{E}[\hat{w}])^2\right]$$

# What are good estimates / estimators?

*Bias* of an estimate is the difference between the value being estimated, and the expected value of the estimate

$$B(\hat{w}) = E[\hat{w}] - w$$

- An *unbiased* estimator has 0 bias

*Variance* of an estimate is, simply its variance, the value of the squared deviations from the mean estimate

$$var(\hat{w}) = E\left[(\hat{w} - E[\hat{w}])^2\right]$$

We want low bias low variance.
But there is a trade-off: reducing one increases the other. low variance results in high bias.

# Estimating parameters: Bayesian approach

Given the training data $\boldsymbol{x}$, we find the *posterior distribution*

$$p(\boldsymbol{w}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{x})}$$

- The result, posterior, is a distribution over the parameter(s)
- One can get a *point estimate* of $\boldsymbol{w}$, for example, by calculating the expected value of the posteriror
- The posterior distribution also contains the information on the uncertainty of the estimate
- A *prior* distribution required for the estimation

# Estimating parameters: frequentist approach

## Maximum likelihood estimation (MLE)

Given the training data $\boldsymbol{x}$, we find the value of $\boldsymbol{w}$ that maximizes the likelihood

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{x}|\boldsymbol{w})$$

- The likelihood function $\mathcal{L}(\boldsymbol{w}|\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{w})$, is a function of the parameters
- The problem becomes searching for the maximum value of a function
- Note that we cannot make probabilistic statements about $\boldsymbol{w}$
- Uncertainty of the estimate is less straightforward

# A simple example

## definition

Problem: We want to estimate the average number of characters in tweets.

Data: We have two data sets (samples)

small $\mathbf{x} = 87, 101, 88, 45, 138$

- The mean of the sample ($\bar{x}$) is 91.8
- Variance of the sample ($sd^2$) is 1111.7 ($sd = 33.34$)

large $\mathbf{x} = (87, 101, 88, 45, 138, 66, 79, 78, 140, 102)$

- $\bar{x} = 92.4$
- $sd^2 = 876.71$ ($sd = 29.61$)

# A simple example

## the task

- We are interested in the mean of all tweets (a large population)
- We only have samples
- Questions:
  - Given a sample, what is the most likely population mean?
  - How certain is our estimate of the population mean?

# A simple example

## the model

$$y = \mu + \epsilon \quad \text{where} \mu \sim \mathcal{N}(0, \sigma^2)$$

Equivalently,

$$y \sim \mathcal{N}(\mu + \sigma^2)$$

- The model is known as the mean/constant/intercept model
- It is related to well-known statistical tests such as t-test (we won't cover it here)

We are normally interested in *conditional models*, models with predictors.

# A simple example

## Bayesian estimation / inference

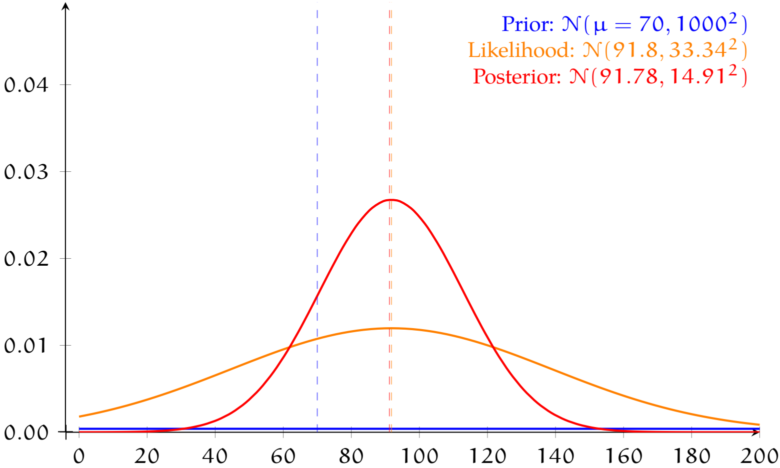
We simply use the Bayes' formula:

$$p(\mu|\mathbf{x}) = \frac{p(\mathbf{x}|\mu)p(\mu)}{p(\mathbf{x})}$$

- With a vague prior (high variance/entropy), the posterior mean is (almost) the same as the mean of the data
- With a prior with lower variance, posterior is between the prior and the data mean
- Posterior variance indicates the uncertainty of our estimate. With more data, we get a more certain estimate
- With a normal prior, posterior will also be normal, and can be calculated analytically

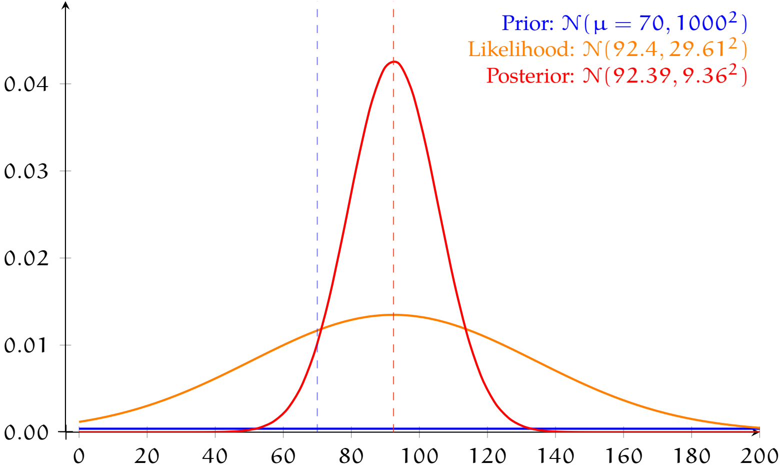
# A simple example

## Bayesian estimation: vague prior, small sample



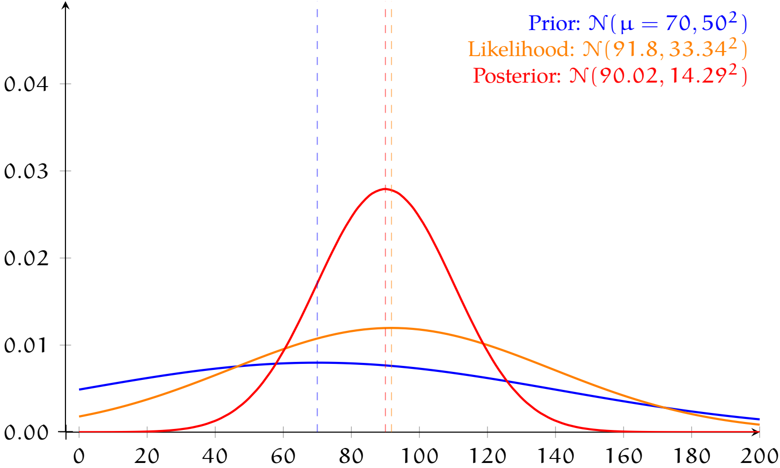
# A simple example

## Bayesian estimation: vague prior, larger sample



# A simple example

## Bayesian estimation: stronger prior, small sample



# A simple example

## MLE estimation

$$
\begin{aligned}
\hat{\mu} &= \arg\max_{\mu} \mathcal{L}(\mu; \boldsymbol{x}) \\
&= \arg\max_{\mu} p(\boldsymbol{x}|\mu) \\
&= \arg\max_{\mu} \prod_{x \in \boldsymbol{x}} p(x|\mu) \\
&= \arg\max_{\mu} \prod_{x \in \boldsymbol{x}} \frac{e^{-\frac{(x-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \\
&= \bar{x}
\end{aligned}
$$

- For 5-tweet sample: $\hat{\mu} = \bar{x} = 91.8$ (cf. 91.78)
- For 10-tweet sample: $\hat{\mu} = \bar{x} = 92.4$ (cf. 92.39)

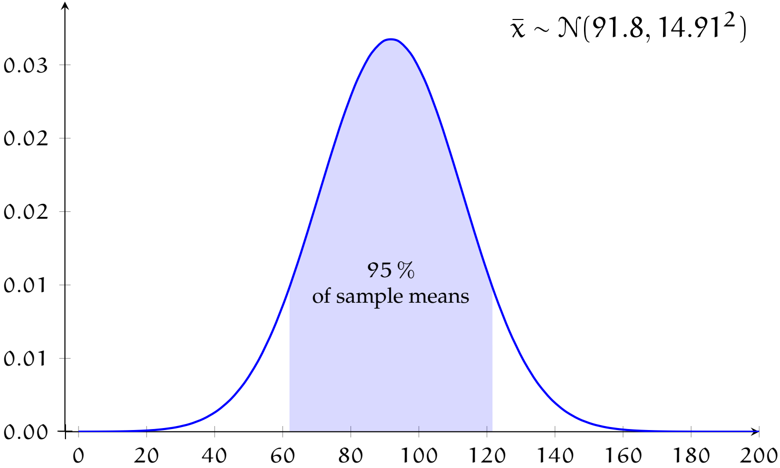
# Classical (frequentist) inference

- We express the uncertainty in terms of the sampling distribution
- Central limit theorem says that means of the samples of size $n$ has a standard deviation of

$$SE_{\bar{x}} = \frac{sd_x}{\sqrt{n}}$$

  - For 5-tweet sample: $SE_{\bar{x}} = 33.34/\sqrt{5} = 14.91$
  - For 10-tweet sample: $SE_{\bar{x}} = 29.61/\sqrt{10} = 9.36$
- A rough estimate for a 95% confidence interval is $\bar{x} \pm 2SE_{\bar{x}}$
  - For 5-tweet sample: $91.8 \pm 2 \times 14.91 = [61.98, 121.62]$
  - For 10-tweet sample: $92.4 \pm 2 \times 9.36 = [83.04, 101.76]$

# Confidence intervals



# Summary / concluding remarks

- Statistical models are important tools in statistical analysis, and machine learning
- There are two major approaches to estimation and inference

Bayesian approach admits a prior distribution, and uses probability theory for inference

Frequentist approach emphasizes unbiased estimates (often MLE), the inference is based on sampling distribution

- The results often agree, but not necessarily

# Next

| | |
|---|---|
| Wed | N-gram language models (1) |
| Fri | Exercises |
| Mon | ML intro: regression and logistic regression |

# Further reading / references

Box, George E. P. and Norman R. Draper (1986). *Empirical Model-Building and Response Surfaces*. New York, USA: John Wiley & Sons, Inc. ISBN: 0-471-81033-9.