L101: Machine Learning for Language Processing

Lecture 1

Guy Emerson

About the course

- Introduction to using Machine Learning for Natural Language Processing

About the course

- Introduction to using Machine Learning for Natural Language Processing
- Prerequisites:
  - L90 (or similar) – essential
  - L95 – desirable

About the course

- Introduction to using Machine Learning for Natural Language Processing
- Prerequisites:
  - L90 (or similar) – essential
  - L95 – desirable
- 8 lectures, 8 seminars, 1 essay/project

Sources of Information

- Course web pages
  - Handouts include additional notes!
- L90 (and L95) notes
- Textbooks, e.g. Jurafsky & Martin

Sources of Information

- Course web pages
  - Handouts include additional notes!
- L90 (and L95) notes
- Textbooks, e.g. Jurafsky & Martin
- Ask questions!

Today's Lecture

- What is Machine Learning?
- Example: topic classification
- How do we know if it works?

What is Machine Learning?

- Task
- Data
- Model
- Training

Tasks

- What do we want to do?

Tasks

- What do we want to do?
- Abstract from a real-world problem

Tasks

- What do we want to do?
- Abstract from a real-world problem
- Examples:
  - Sentiment analysis
  - Topic classification
  - Machine translation

Data

- Types of data:
  - Natural (e.g. “raw” text)
  - Pre-processed (e.g. tokenised text)
  - Annotated (e.g. pos-tagged text)

Supervised vs. Unsupervised

$$f : x \mapsto y$$

Supervised vs. Unsupervised

$$f : x \mapsto y$$

input output

Supervised vs. Unsupervised

$$f : x \mapsto y$$

- Supervised: observe pairs (x, y)

Supervised vs. Unsupervised

$$f : x \mapsto y$$

- Supervised: observe pairs (x, y)
- Unsupervised: observe only x

Supervised vs. Unsupervised

$$f : x \mapsto y$$

- Supervised: observe pairs (x, y)
- Unsupervised: observe only x
- Semi-supervised: observe both

Models

$$f : x \mapsto y$$

- How do we represent f?

Models

$$f : x \mapsto y$$

- How do we represent f?
- Parameters

Discriminative vs. Generative

$$f : x \mapsto y$$

- Non-probabilistic: f
- Discriminative: $P(y|x)$
- Generative: $P(x, y)$

What is Machine Learning?

- Task
- Data
- Model
- Training

What is Machine Learning?

- Task – what function do we want?
- Data
- Model
- Training

What is Machine Learning?

- Task – what function do we want?
- Data – what do we observe?
- Model
- Training

What is Machine Learning?

- Task – what function do we want?
- Data – what do we observe?
- Model – how do we represent the function?
- Training

What is Machine Learning?

- Task – what function do we want?
- Data – what do we observe?
- Model – how do we represent the function?
- Training – how do we fix the representation, based on what we observe?

Topic Classification

- Task
  - Input: text
  - Output: topic (out of small set)

Topic Classification

- Task
  - Input: text
  - Output: topic (out of small set)
- Data
  - Texts, each labelled with a topic

Topic Classification

- Task
  - Input: text
  - Output: topic (out of small set)
- Data
  - Texts, each labelled with a topic
  - (If unsupervised: topic *discovery*)

Naive Bayes

- Generative model

Naive Bayes

- Generative model – $P(x, y)$

Naive Bayes

$$\underset{y}{\text{argmax}}\, P(y|x)$$

Naive Bayes

$$\underset{y}{\text{argmax}}\, P(y|x) \;=\; \underset{y}{\text{argmax}}\, P(y)P(x|y)$$

Naive Bayes

$$\begin{aligned}
\operatorname*{argmax}_y P(y|x) &= \operatorname*{argmax}_y P(y)P(x|y) \\
&\approx \operatorname*{argmax}_y P(y)\prod_i P(x_i|y)
\end{aligned}$$

Naive Bayes

Bayes

$$\underset{y}{\text{argmax}}\, P(y|x) = \underset{y}{\text{argmax}}\, P(y)P(x|y)$$

$$\approx \underset{y}{\text{argmax}}\, P(y)\prod_i P(x_i|y)$$

Naive Bayes

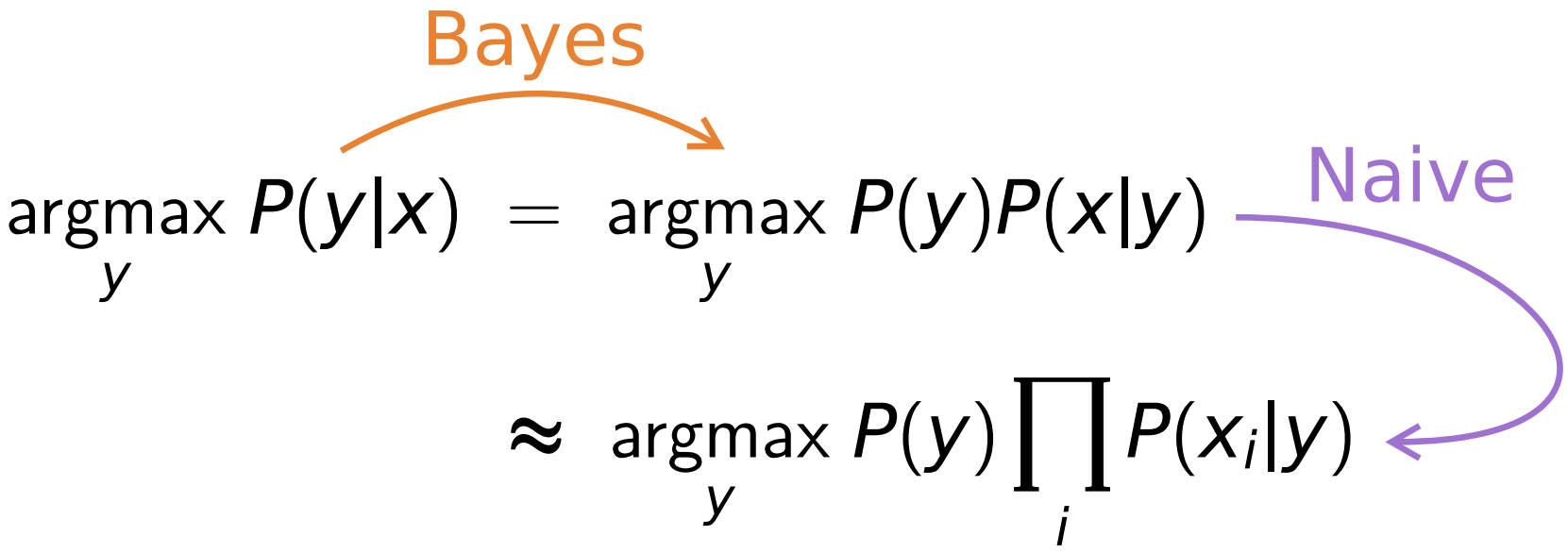
Bayes

$$\underset{y}{\text{argmax}}\, P(y|x) = \underset{y}{\text{argmax}}\, P(y)P(x|y)$$

Naive

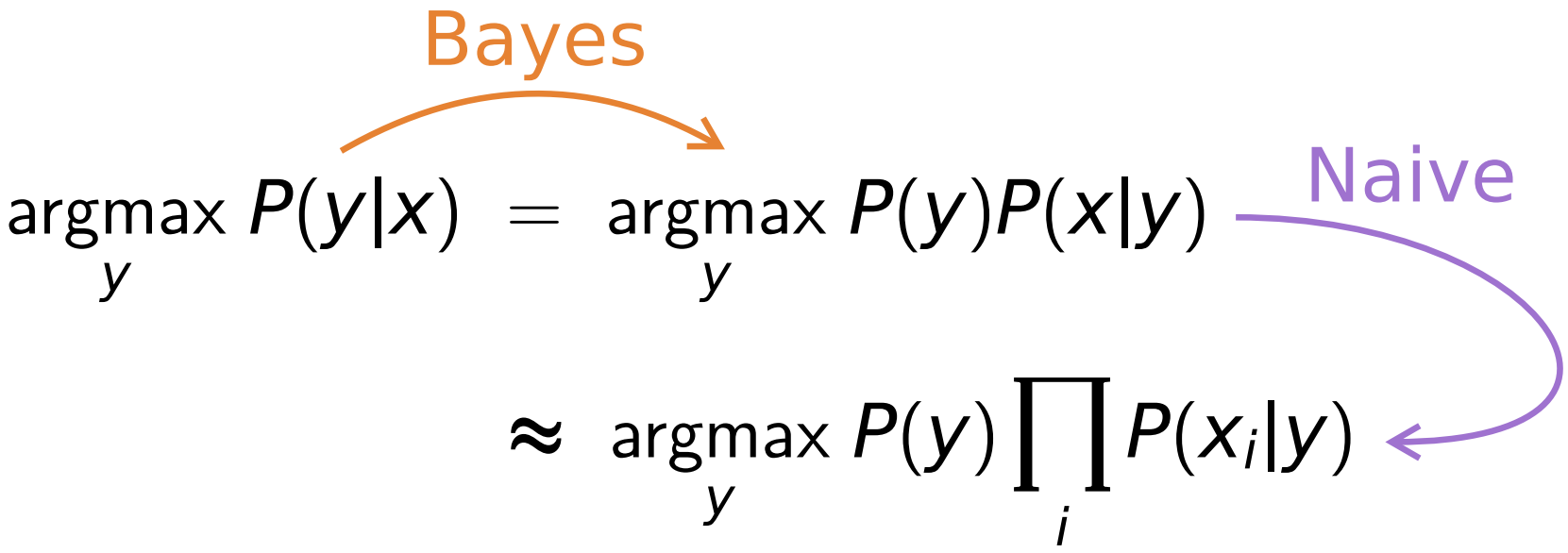
$$\approx \underset{y}{\text{argmax}}\, P(y)\prod_i P(x_i|y)$$

Naive Bayes



- Bernoulli NB – x_i binary-valued

Naive Bayes



- Bernoulli NB – x_i binary-valued
- Multinomial NB – x_i integer-valued

Naive Bayes

- Parameters: $P(y)$, $P(x_i|y)$

Naive Bayes

- Parameters: $P(y)$, $P(x_i|y)$
- Training (Bernoulli NB):
  - $P(y) = \frac{N_y}{N}$
  - $P(x_i|y) = \frac{N_{y,i}}{N_y}$

Naive Bayes

- Parameters: $P(y)$, $P(x_i|y)$
- Training (Bernoulli NB):
  - $P(y) = \frac{N_y + \alpha}{N + K\alpha}$
  - $P(x_i|y) = \frac{N_{y,i} + \beta}{N_y + 2\beta}$

Naive Bayes

- Parameters: $P(y)$, $P(x_i|y)$
- Training (Bernoulli NB):
  - $P(y) = \frac{N_y + \alpha}{N + K\alpha}$
  - $P(x_i|y) = \frac{N_{y,i} + \beta}{N_y + 2\beta}$
- Hyperparameters: α, β

Example: English Wikipedia

Thus, what started as an effort to translate between languages evolved into an entire discipline devoted to understanding how to represent and process natural languages using computers.

Example: English Wikipedia

An extreme example is the alien species, the Vulcans, who had a violent past but learned to control their emotions.

Example: German Wikipedia

Es umschließt die Mündungen des Hudson River und des East River in den Atlantischen Ozean und erhebt sich durchschnittlich sechs Meter über den Meeresspiegel.

Example: German Wikipedia

Es umschließt die Mündungen des Hudson River und des East River in den Atlantischen Ozean und erhebt sich durchschnittlich sechs Meter über den Meeresspiegel.

Example: German Wikipedia

Schließlich bediente sich Ian Fleming auch der Geschichten und des Charakters des serbischen Doppelagenten Duško Popov aus dem Zweiten Weltkrieg.

Example: German Wikipedia

Schließlich bediente sich Ian Fleming auch der Geschichten und des Charakters des serbischen Doppelagenten Duško Popov aus dem Zweiten Weltkrieg.

Evaluation

How do we know if it works?

Training and Testing

- Split data:
  - Training
  - Development
  - Testing

Training and Testing

- Split data:
  - Training
  - Development
  - Testing
- Metric (e.g. accuracy, F1)

Training and Testing

- Split data:
  - Training
  - Development
  - Testing
- Metric (e.g. accuracy, F1)
- Baseline, significance test

Shared Tasks

- Task
- Data

Provided

- Model
- Training

Participant

Summary

- ML – task, data, model, training
- Topic classification with Naive Bayes
- Evaluation – data split, shared tasks