Naïve Bayes Classifier

Pattern Recognition 2018

Adam Kortylewski

University of Basel

Text classification

- What is the subject category, topic or genre of an article?

IMDB Genre Classification – R-bloggers, Florian Teschner

Text classification

- What is the subject category, topic or genre of an article?

- Positive or negative movie review?


- unbelievably disappointing


- Full of zany characters and richly applied satire, and some great plot twists


- this is the greatest screwball comedy ever filmed


- It was pathetic. The worst part about it was the boxing scenes.

Text classification – Dan Jurafsky

Text classification

- What is the subject category, topic or genre of an article?
- Positive or negative movie review?
- Spam Detection
- Authorship identification
- …

Text classification is difficult

- Text is highly structured data
  - The context can have a strong influence on the meaning of a word
- Topics can vary wildly
- Misclassification cost may be asymmetric
  - Spam detection: We do not want to lose important messages and we do not mind disposing of a few individual spam messages manually.

Spam classification

Received: from mx1-pub.urz.unibas.ch (131.152.226.162) by exch.unibas.ch (131.152.8.132) with Microsoft SMTP Server id 14.3.174.1; Wed, 28 May 2014 12:21:57 +0200
From: "bis zum 8. Juni" xbnmmsjgnscfh@gareau.toyota.ca
To: mifdav00@stud.unibas.ch
Subject: Ruby Palace Handycasino verdreifacht heute Ihre Einzahlung

Hallo,

Sie haben Glück! Ihr und unser guter Freund, Christian, hat eine Glückssträhne bei uns im Ruby Palace Casino – er gewann £/$/€640 auf Blackjack – und nun möchte er, dass Sie es ihm gleichtun und in den Gewinnerkreis einsteigen.

Ruby Palace bietet Ihnen nur das Beste – von einer großartigen Auszahlungsrate von 97 Prozent bis hin zur einer exklusiven Auswahl an spannenden Spielen, einschließlich Spieltischen sowie beliebte Spielautomaten und vieles mehr. Zudem steht Ruby Palace für fairen Spielbetrieb und verantwortungsvolle Casinoführung.

Als ein Freund von Christian, und er hat dich mit Begeisterung empfohlen, erhalten Sie ein Willkommensgeschenk von 200% auf Ihre erste Einzahlung, wenn Sie sich noch heute anmelden.

Beginnen Sie noch heute! Sagen Sie "Ja" und melden Sie sich heute an.

Viel Glück!

CALL FOR PARTICIPATION

The organizers of the 11th IEEE International Conference on Automatic Face and Gesture Recognition (IEEE FG 2015) **invite interested research groups to participate in the special sessions and workshops** organized as part of IEEE FG 2015. Accepted papers will be published as part of the **Proceedings of IEEE FG2015 & Workshops** and submitted for inclusion into IEEE Xplore.

Special sessions (http://www.fg2015.org/participate/special-sessions/):

1. ANALYSIS OF MOUTH MOTION FOR SPEECH RECOGNITION AND SPEAKER VERIFICATION

Organizers: Ziheng Zhou, Guoying Zhao, Stefanos Zafeiriou

Submission deadline: 24 November, 2014

2. FACE AND GESTURE RECOGNITION IN FORENSICS

Organizers: Julian Fierrez, Peter K. Larsen, (co-orgnized by COST Action IC 1106)

Submission deadline: 24 November, 2014

Recap: General Bayes Classifier

- For a document x and a class c:

$$\begin{aligned} c^* &= \arg\max_c P(c|\vec{x}) \\ &= \arg\max_c \frac{P(\vec{x}|c)P(c)}{P(\vec{x})} \\ &= \arg\max_c P(\vec{x}|c)P(c) \end{aligned}$$

- What is the representation $\vec{x}$ of a text document?

Naïve assumption (1) – Bag of words representation

Received: from mx1-pub.urz.unibas.ch (131.152.226.162) by exch.unibas.ch (131.152.8.132) with Microsoft SMTP Server id 14.3.174.1; Wed, 28 May 2014 12:21:57 +0200
From: "bis zum 8. Juni" xbnmmsjgnscfh@gareau.toyota.ca
To: mifdav00@stud.unibas.ch
Subject: Ruby Palace Handycasino verdreifacht heute Ihre Einzahlung

Auszahlungsrate	3
Glückssträhne	2
Geschenk	2
Spieltischen	1
Geld	5
…	

Begeisterung empfohlen, erhalten Sie ein Willkommensgeschenk von 200% auf Ihre erste Einzahlung, wenn Sie sich noch heute anmelden.

Beginnen Sie noch heute! Sagen Sie "Ja" und melden Sie sich heute an.

Viel Glück!

CALL FOR PARTICIPATION

The organizers of the 11th IEEE International Conference on Automatic Face and Gesture Recognition (IEEE FG 2015) **invite interested research groups to participate in the special sessions and workshops** organized as part of IEEE FG 2015. Accepted papers will be published

Research	2
Proceedings	5
Recognition	2
Face	3
Submission	1
…	

Submission deadline: 24 November, 2014

2. FACE AND GESTURE RECOGNITION IN FORENSICS

Organizers: Julian Fierrez, Peter K. Larsen, (co-orgnized by COST Action IC 1106)

Submission deadline: 24 November, 2014

The order of the words is lost!

Recap: General Bayes Classifier

- For a document x and a class c:

$$\begin{aligned} c^* &= \arg\max_c P(c|\vec{x}) \\ &= \arg\max_c \frac{P(\vec{x}|c)P(c)}{P(\vec{x})} \\ &= \arg\max_c P(\vec{x}|c)P(c) \end{aligned}$$

- It is difficult to estimate $P(\vec{x}|c) = P(x_1, x_2, x_3, \dots, x_M|c)$
  - Enormous amounts of parameters needed
  - Missing data

Naïve assumption (2) – Conditional independence

- (1) Bag of words representation – assume position does not matter

- (2) Conditional independence – assume feature probabilities are independent given the class

$$P(\vec{x}|c) = P(x_1, x_2, x_3, \dots, x_M|c)$$

$$= P(x_1|c)P(x_2|c), \dots, P(x_M|c)$$

General Naïve Bayes Classifier

The Naïve Bayes classifier is more than a spam or text classifier. It is a general classification model based on the Bayes classifier with an additional assumption: *Conditionally independent features*.

- Probabilistic Classifier (is a Bayes classifier)
  - E.g. for asymmetric loss function
- Completely factorized using *"conditional independence"*
  - Independent features conditional on the class: efficient and simple
- Features can have their own distribution each
  - Even continuous and discrete distributions mixed
- Generative: handles missing data and unlabeled data (EM or alike)

Naïve Bayes Classifier for text classification

- Message email is a collection of independent words w

$$P(\text{email}|c) = P(c) \prod_{w \in \text{email}} P(w|c) \qquad c = \begin{cases} \text{ham} \\ \text{spam} \end{cases}$$

$$\sum_{w} P(w|c) = 1$$

- Each word is drawn from a vocabulary with probability $P(w|c)$

  Occurrence in vocabulary is specific to each class

- Parameter estimation: Maximum Likelihood

Parameter Estimation

$$p(c) = \frac{N_c}{\sum_{c'} N_{c'}}$$

Relative frequency of the document class c in the training set

$$p(w|c) = \frac{N_{wc}}{\sum_{w'} N_{w'c}}$$

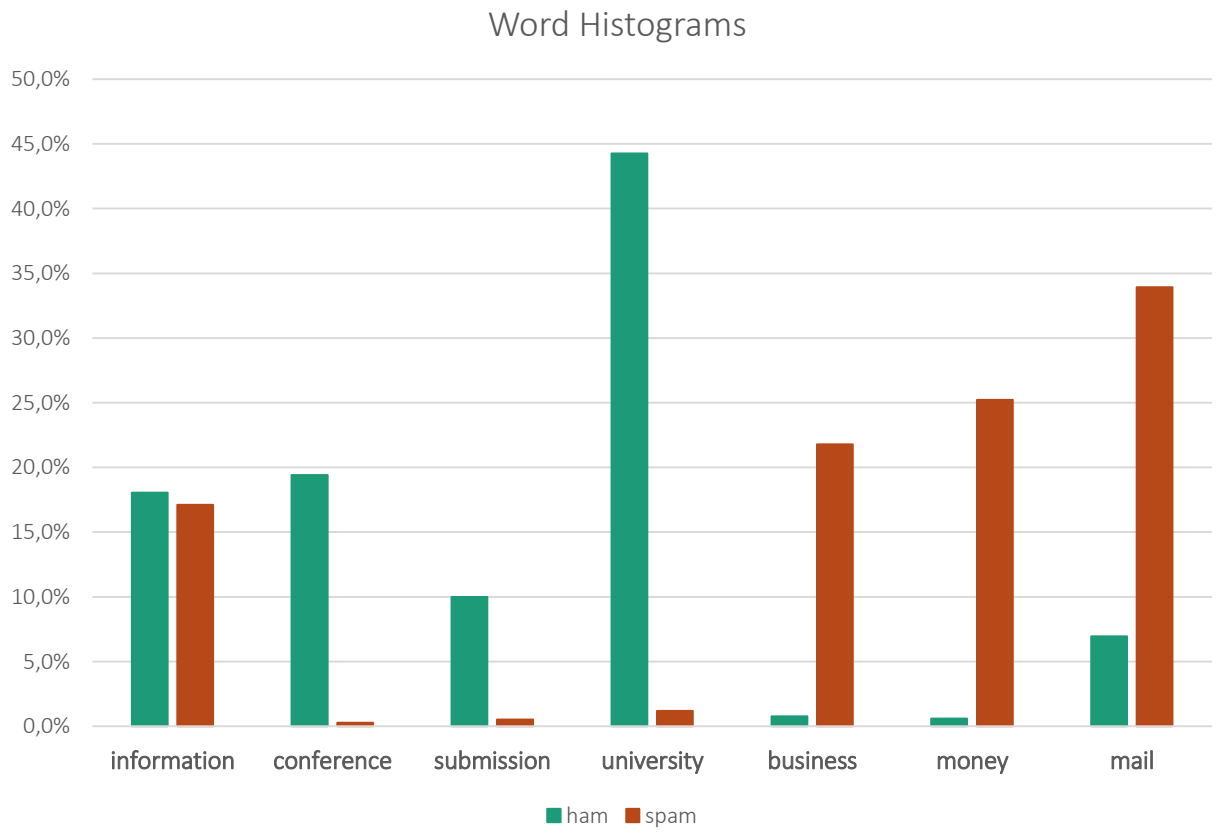
Relative frequency of a word w in the training set

- What if a word does not occur in a document class?

$$p(w|c) = \frac{(N_{wc}+1)}{\sum_{w'}(N_{w'c}+1)}$$

Laplace smoothing for Naïve Bayes

Bag-of-Words Model: Word Histograms



word	P(w|ham)	P(w|spam)
information	18.0%	17.1%
conference	19.4%	0.3%
submission	10.0%	0.5%
university	44.3%	1.2%
business	0.8%	21.8%
money	0.6%	25.2%
mail	6.9%	33.9%

“vocabulary”

Bag-of-Words Model: Classification

- Classification rule: find best class c^* of message m

$$c^* = \arg\max_c P(c|m)$$

Largest posterior: Bayes classifier

$$c^* = \arg\max_c P(c) \prod_{w \in m} P(w|c)$$

Independent words

All words in message (in dictionary)

$$c^* = \arg\max_c \log P(c) + \sum_{w \in m} \log P(w|c)$$

log: numerical accuracy

Careful: If you work with explicit counts and you need to compare different messages with each other, you need an additional normalization factor which depends on the message length (*multinomial distribution*)

Bag-of-Words Model: Scoring

- Log of posterior ratio can be interpreted as a spam score

- Each word *adds* to the total spam score of the message

$$r = \log \frac{P(\mathrm{s}|m)}{P(\mathrm{h}|m)} = \log \frac{P(m|\mathrm{s})P(\mathrm{s})}{P(m|\mathrm{h})P(\mathrm{h})} = \log \frac{P(\mathrm{s})}{P(\mathrm{h})} + \sum_{w \in m} \log \frac{p(w|s)}{p(w|h)}$$

Email Spam Detection

- Email messages as input

  Whole message, including headers, as pure text

- Preprocessing (training set and test emails)
  - Split into words: *tokenization*
  - Remove stop words: *and, of , if, or, …* (optional)
  - Stemming: replace word forms by a common class (optional)
    e.g. *includes, included, include, …*
- Learning: word counts → word likelihoods
- Classification: scoring with likelihoods of words in message

Vocabulary Reduction

Removing words from the vocabulary

- Dealing with words which are unseen in training data
- Reducing classifier to most important words only: Optimization

- What is important/significant?

  High or low spam score: word contains information relevant for classification

- During classification, ignore words which are not in the dictionary

Word Counting: Likelihood Models

Our word counting heuristic is a sound probabilistic likelihood model

- Multinomial distribution:

$$P(N_1, N_2, \dots, N_W | \, c) = \frac{(N_1 + N_2 + \cdots + N_W)!}{N_1!\, N_2! \cdots N_W!} p(w_1|c)^{N_1} p(w_2|c)^{N_2} \cdots p(w_W|c)^{Nw}$$

Probability of absolute word frequencies N_w for W words

Occurrence probabilities $p(w|c)$

We can also use a different likelihood model in Naïve Bayes:

- Binomial distribution: A word occurs or does not (Boolean)
  - Does not care how many times a word appears
  - Missing words also appear in likelihood term

$$P(\theta_1, \theta_2, \dots, \theta_W | \, c) = \prod_w p(w_1|c)^{\theta_w} \big(1 - p(w_1|c)\big)^{1-\theta_w} \quad \theta_w \in \{0, 1\}$$

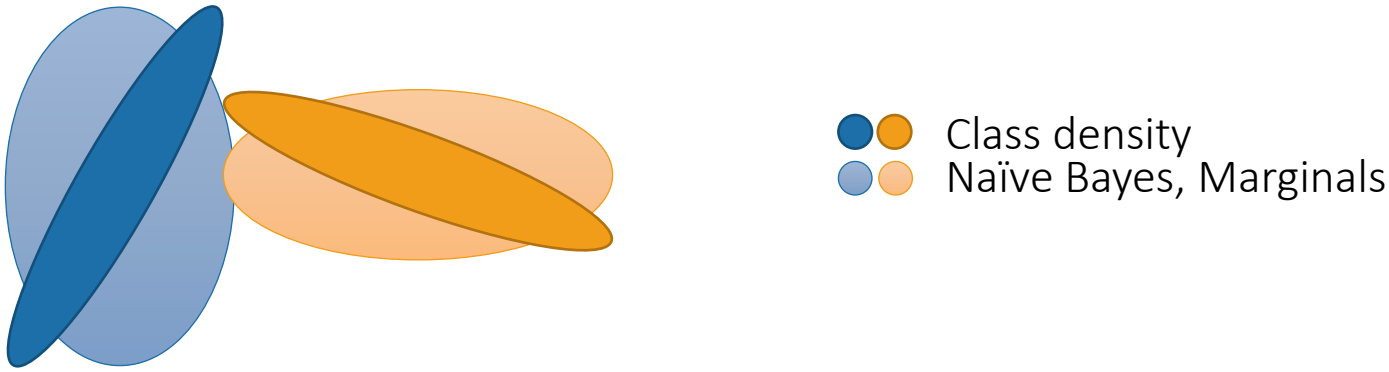
$\theta_w = 1$: word occurs

Conditionally Independent Features

- Conditional independence assumption is rarely appropriate

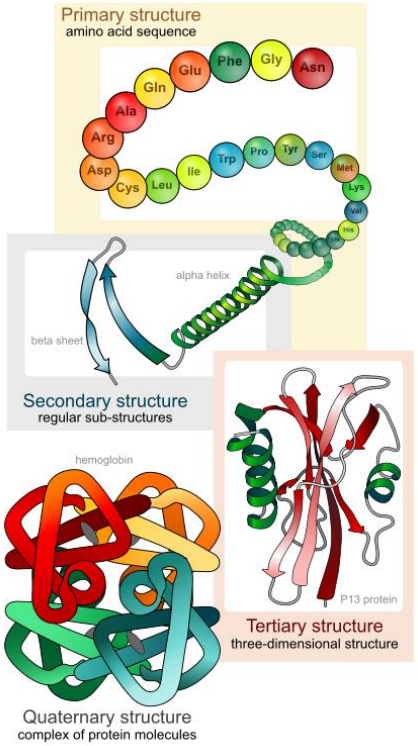
  It ignores correlation between features of a class. But features in the same class are very often highly correlated. For example the size and weight of a fish, even in a single class.

- Bad representation of the class density:
  product of feature *marginals* only, no correlation

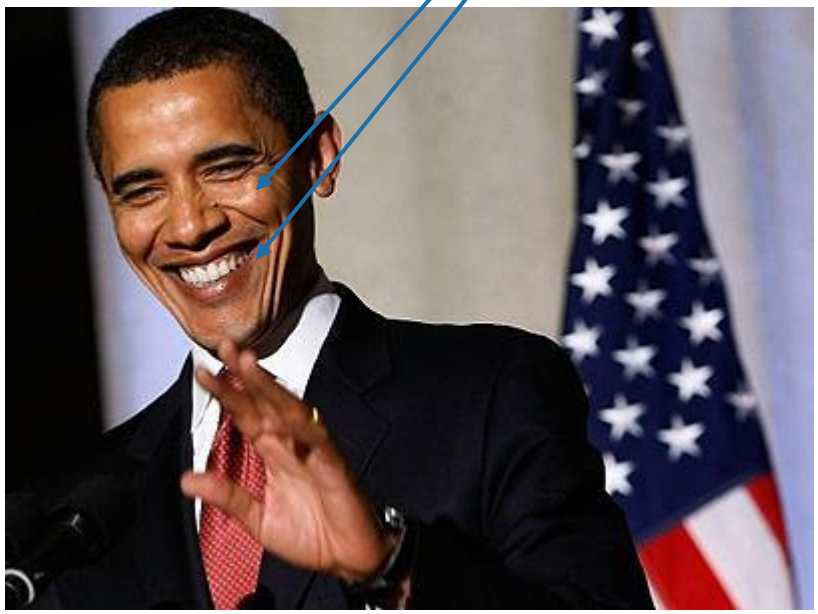


Data is usually Heavily Structured

Structure (=dependence) is very important in real-word situations



Relations among pixels (dependencies)



Genetic Code → Function

Image → Facial Expression

Factorization: Level of Independence

The likelihood is optimally the *joint distribution* of all features $\vec{x}$ in class c_i: $P(\vec{x}|c_i)$

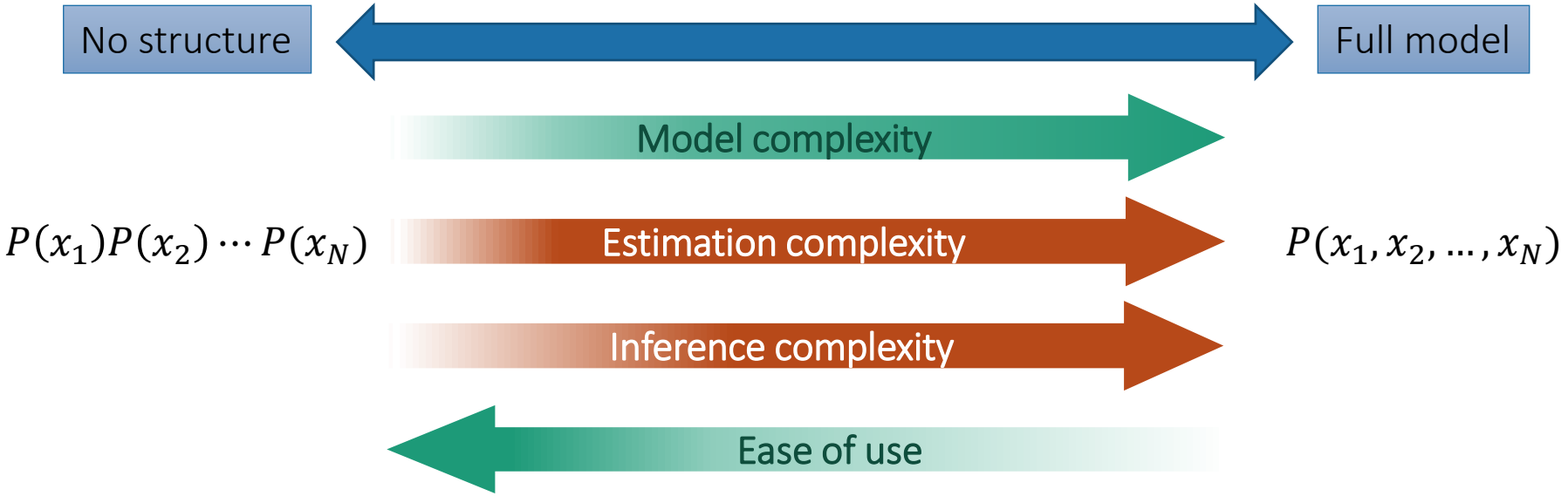
- Very difficult to obtain in high-dimensional space

  e.g. whole images: 1MP → 10^6 dimensional features!

- Naïve Bayes: loss of all structure – complete independence in each class – can lead to good results. Is structure unnecessary?

There are *probabilistic graphical models*: A formal method to model the structure of the problem – specify which *links* between features should be kept and estimated. It is possible to capture expert knowledge about a domain and use it for classification.

More Structure?

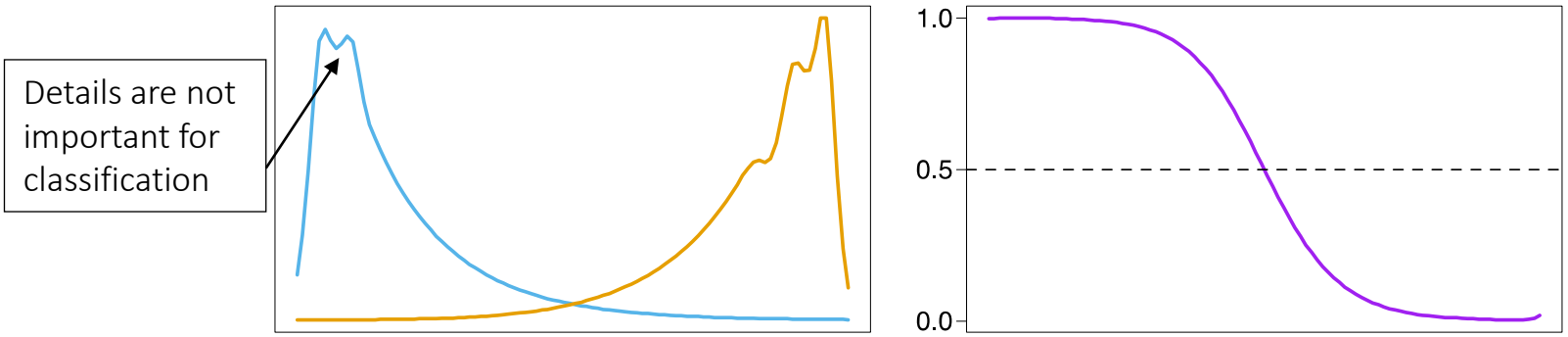


Can we characterize a middle way? – Yes: *Graphical Models*

- Use just *some* structure: experts! – Explicit knowledge integration
- Tradeoff between complexity and applicability
- Representation of intermediate levels of structure?

Does Independence Hurt?

- Posterior of Naïve Bayes can still be very good, even with a bad class representation – often generative modeling invests where it is not important for classification
- Estimation of marginals is so much easier than full joint estimation: With limited data, the "incorrect" model often outperforms the "real" one, especially in high-dimensional spaces.



Class densities (1D) Posterior (blue class)

Generative vs. Discriminative

We want to classify based on the posterior distribution

$$P(c_i|\vec{x})$$

It can be modeled in conceptually different ways:

- Generative Model: Known Bayes classifier

  Likelihood and prior models form the posterior using Bayes rule

  $$P(c_i|\vec{x}) \propto p(\vec{x}|c_i)P(c_i)$$

- Discriminative Model:

  Directly estimate the posterior distribution $P(c_i|\vec{x})$

  Known *algorithmically*: SVM, Perceptron, Logistic regression, etc.

Generative vs. Discriminative (II)

- Naïve Bayes is generative
- Generative models have benefits
  - Model can generate artificial samples: model sanity
  - Deal with missing data & hidden variables (~EM)
  - Expert knowledge guides structure
  - Extensible: can add new factors if necessary without invalidating the model
- Generative models have a big disadvantage
  - Waste modeling effort where it might not be important

Summary: Naïve Bayes Classifier

- Bayes classifier with the assumption of independent features
  - Probabilistic, generative classifier
  - Easy-to-estimate likelihoods: Product of feature marginals

$$P(x_1, x_2, x_3, \dots, x_M|c) = P(x_1|c)P(x_2|c)P(x_3|c)\cdots P(x_M|c)$$

  - Can deal with different distributions for each feature
- Application to text classification with the *bag-of-words* model:
  - Content-based classification: Text as a collection of words
  - Order of words is not important
  - Word occurrence histograms (Multinomial likelihood model)
  - Easy classification by summing word scores