CIS501 – Lecture 6

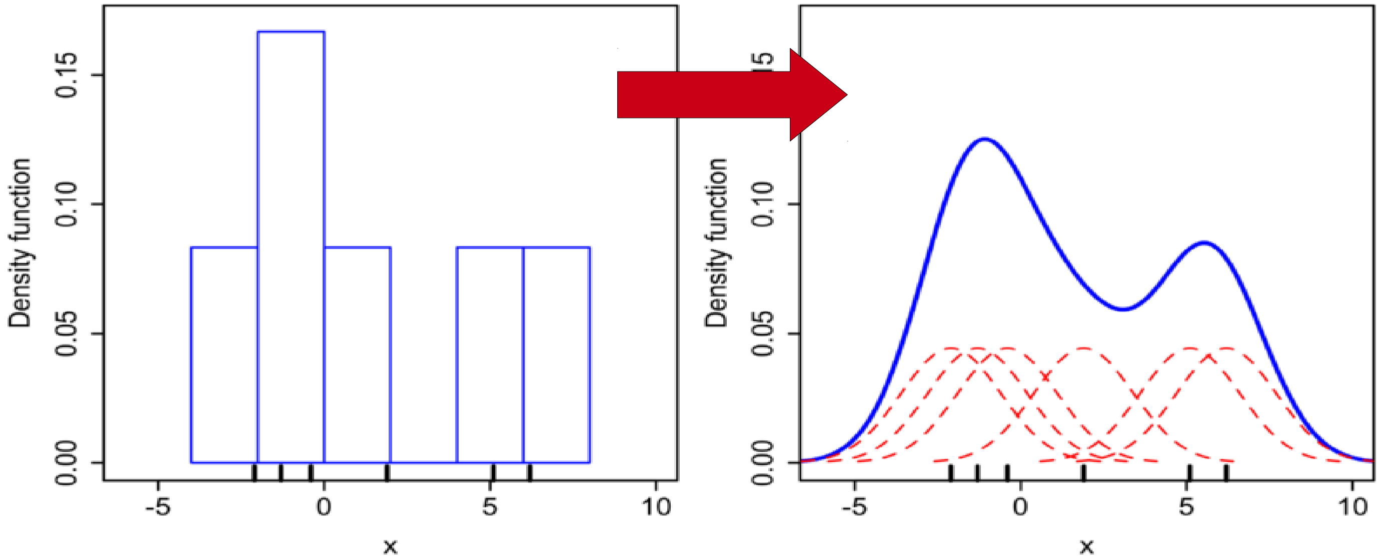
Woon Wei Lee
Fall 2013, 10-11:15am,
Sundays and Wednesdays

Masdar INSTITUTE

For today:

- Administrative stuff
- *k*NN wrap-up
- Naïve Bayes Classifier
  - Spam filtering example
  - Multivariate Bernoulli *vs* Multinomial event models
- Presentations:
  - Timothy Mulumba
  - Andor Kovacs

k-NN as a form of kernel density estimation

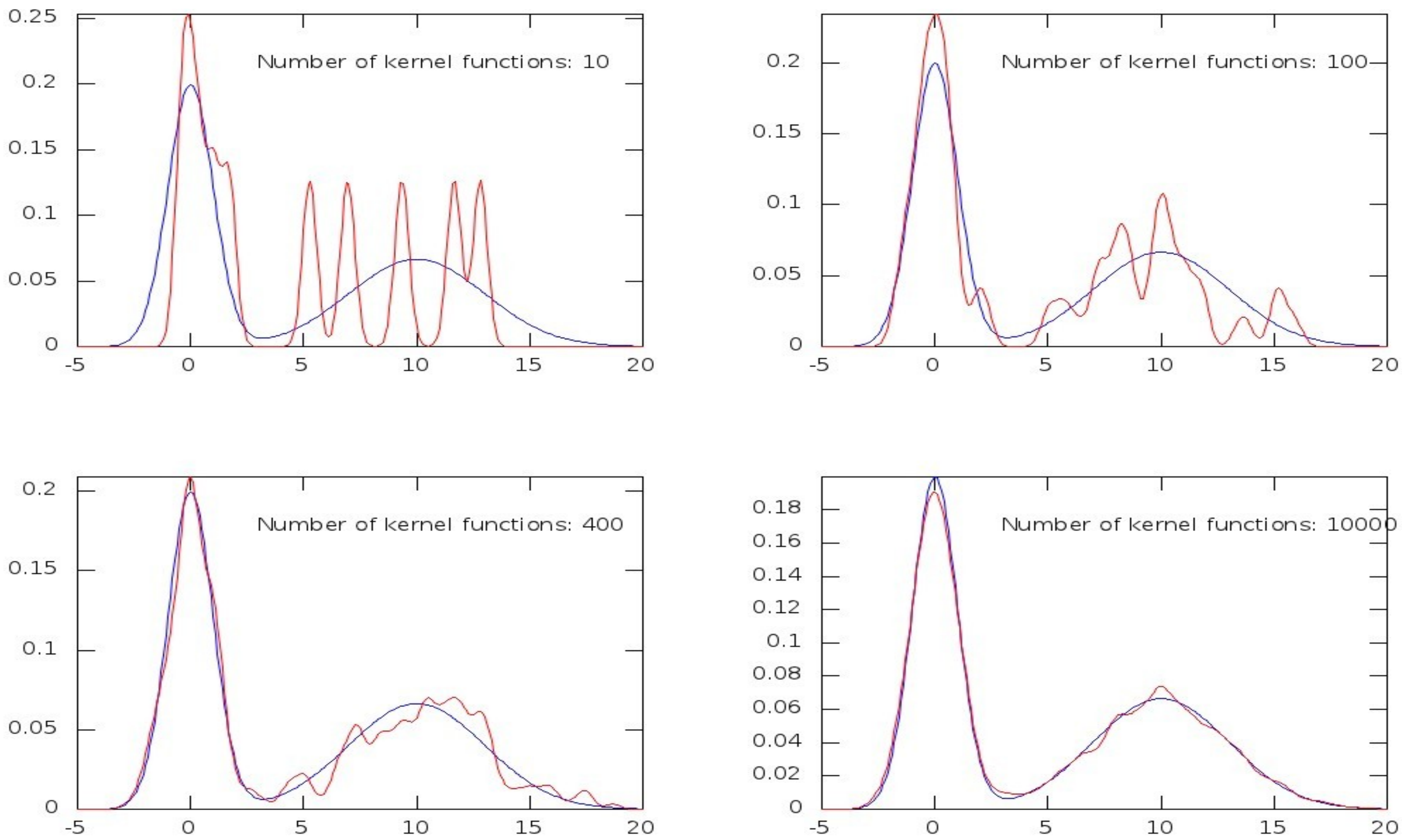


- Standard (unweighted) k-NN assumes that each of the k closest points contributes a uniform probability density to p(x|c).
- The distance weighted k-NN assumes a unimodal density (depending on weighting function).
- Related to the technique of kernel density estimation is a technique where PDF is approximated via:

$$p(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, x_i)$$

- (But, only k-nearest training neighbours considered)
- In the special case where k → number of training points, then kNN is exactly kernel density estimation.

Cont'd



- **Example of kernel density estimation.**
  - “True density” → mixture of two gaussians, N(0,1) and N(10,3)
  - Kernels → gaussians with std of 0.1
- **Increase in number of kernel functions → greater smoothness**
- **Problem → high dimensional spaces...**

Naïve Bayes Classifier

- Another non-parametric, generative classifier
- For a multivariate variable $\boldsymbol{x} \sim \{x_1, x_2, x_3 \ldots x_n\}$:

$$P(c_i|x) = \frac{p(x|c_i)\,p(c_i)}{p(x)}$$

$$= \frac{p(x_1, x_2, \cdots x_n|c_i)\,p(c_i)}{p(x)}$$

- As before, we want to build a classifier based on the Bayes decision rule..
- Problem: How do we estimate the joint probability *p(x1,x2...xn|c_i)*?

Masdar INSTITUTE

NBC (Cont'd)

- The NBC solves this by assuming that the class conditional distributions are all independent.
- Two variables *x* and *y* are independent if:

  p(x,y)=p(x)p(y)

- Similarly:

$$P(c_i|x) = \frac{p(x_1, x_2, \cdots x_n|c_i)\, p(c_i)}{p(x)}$$

$$= \frac{p(x_1|c_i) \cdot p(x_2|c_i) \cdots p(x_n|c_i)\, p(c_i)}{p(x)}$$

- For this reason, NBC is sometimes called the *independent feature model*
- Extremely simple yet still very effective – used extensively for Spam filtering

NBC – Spam filtering example

- Let's look at a common classification task – spam filtering!
- Data model → data comes in word vectors:

$$\mathcal{D} = [\{w_1^1, w_2^1, w_3^1, \ldots, w_{n-1}^1, w_n^1\}, \{w_1^2, w_2^2, w_3^2, \ldots, w_{n-1}^2, w_n^2\}]$$

⇒ Challenge is to classify a previously unseen e-mail:

$$\mathcal{D}_i = \{w_1^i, w_2^i, w_3^i, \ldots, w_{n-1}^i, w_n^i\}$$

into one of two classes → {Spam (c_1), Non-Spam (c_2)}

- By Bayes Theorem: $$p(c|D) = p(c|w_1, w_2, \cdots, w_n) = \frac{p(w_1, w_2, \cdots, w_n|c)\, p(c)}{p(D)}$$

- Calculating the joint likelihood term, $p(w_1, w_2, \ldots, w_n|c)$ is normally very difficult → use the NBC assumption to simplify!

Masdar INSTITUTE

Spam filtering example (Cont'd)

- Assuming inter-word independence, we get:

$$p(w_1, w_2, \cdots, w_n|c) = p(w_1|c).\ p(w_2|c).\cdots p(w_n|c)$$

- The individual $p(w_i|c)$'s are easily estimated from the training set by counting the occurrence frequencies for each respective class.
- Let's assume that our vocabulary includes 3 words:
  - "Viagra" (w_1) appears in 25 out of 30 spam e-mails, so we set → $p(w_1|c_1)=5/6$
  - *"Account"* (w_2) appears in 20 out of 30 spam e-mails, so we set → $p(w_2|c_1)=2/3$
  - *"Password"* (w_3) appears in 18 out of 30 spam e-mails, so we set → $p(w_3|c_1)=3/5$
- Similarly, let's say that $p(w_1|c_2)=1/20$, $p(w_2|c_2)=1/2$, $p(w_3|c_2)=1/2$
- Test e-mail:

  $\mathcal{D}$: "*..natural* ***viagra****! it will… please send us your* ***account****...*"

(Cont'd)

$$\begin{aligned} p(D|c_1) &= p(w_1, w_2, w_3|c_1) \\ &= p(w_1|c_1).\ p(w_2|c_1). \cdots p(w_n|c_1) \\ &= \frac{5}{6} \times \frac{2}{3} \times \frac{2}{5} = \frac{2}{9} \end{aligned}$$

$$\begin{aligned} p(D|c_2) &= p(w_1, w_2, w_3|c_2) \\ &= p(w_1|c_2).\ p(w_2|c_2). \cdots p(w_n|c_2) \\ &= \frac{1}{20} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{80} \end{aligned}$$

- *p(c)* (the *prior*) is simply the relative proportions of each respective class → let's say that $p(c_1)$=1/10 and $p(c_2)$=9/10

  ⇒ $p(c_1|D) \propto 2/9 \times 1/10 = 2/90$

  ⇒ $p(c_2|D) \propto 1/80 \times 9/10 = 9/800$

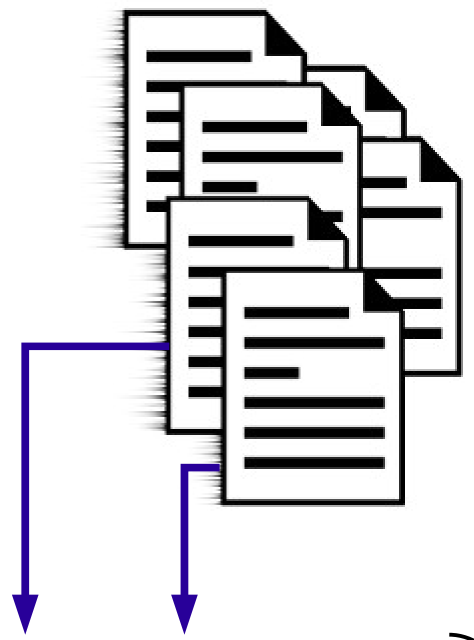
  (p($\mathcal{D}$) (the *evidence*) is class-independent)

- Hence, we could classify this document as a Spam e-mail!
- **Question: How valid do you think the naïveness assumption is in this case?**

Multinomial event model

- The previous example is an instance of the "Multivariate Bernoulli" event model
    - The "canonical" or spreadsheet representation described before
    - Each document is encoded as a vector
    - Sometimes referred to as "***bag-of-words***" model
    - One weakness is that whether a word appears once or one hundred times → final representation is the same!
- An alternative representation is as a "***stream-of-words***"
- Distribution of words is modelled by a *multinomial distribution*

→ "Multinomial Event Model"

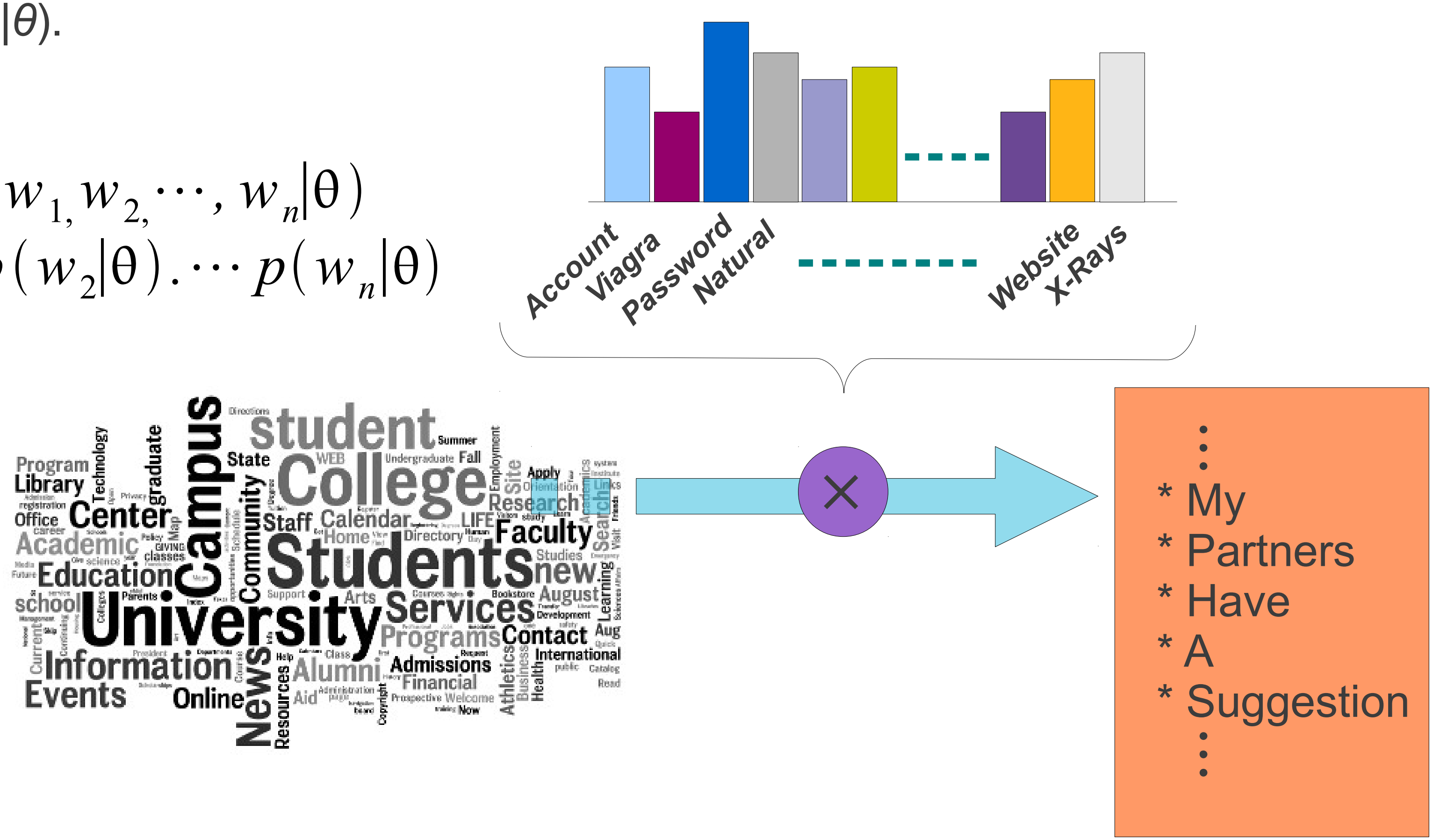


$$\begin{pmatrix} p(w_1^1|c) & p(w_1^2|c) & \\ p(w_2^1|c) & p(w_2^2|c) & \\ p(w_3^1|c) & p(w_3^2|c) & \cdots \\ \vdots & \vdots & \\ p(w_i^1|c) & p(w_i^2|c) & \end{pmatrix}$$

Multinomial event model (Cont'd)

- Characterized by a word "generator" which follows the multinomial distribution
- For a multinomial R.V. θ, each word has its own $p(w_i|\theta)$.
- Hence:

$$p(D|\theta) = p(w_1, w_2, \cdots, w_n|\theta)$$
$$= p(w_1|\theta) . p(w_2|\theta) . \cdots p(w_n|\theta)$$



Comparison: Bernoulli *vs* Multinomial cases

- **Spam example again (sorry ;-))**
- **Multivariate Bernoulli event model:**
  - Vocabulary: {*Viagra, Account, Password*}
  - $p(w_i=1|c_1)=\{5/6,2/3,3/5\}$

    $\rightarrow$ $p(w_i=0|c_1)=\{1/6,1/3,2/5\}$
  - Note that they sum to one for each feature across possible values
  - For the following phrase:

"$\mathcal{D}$: "..*natural* ***viagra****! it will… please send us your* ***account****...*"

$$p(D|c_1)=p(w_1,w_2,w_3|c_1)$$
$$=p(w_1|c_1).\,p(w_2|c_1).\cdots p(w_n|c_1)$$
$$=\frac{5}{6}\times\frac{2}{3}\times\frac{2}{5}=\frac{2}{9}$$

- **Multinomial event model:**
  - Vocabulary: {*Viagra, Account, Password*}
  - $p(w_i|c_1)=\{3/6,1/3,1/6\}$
  - Note that they sum to one across all features
  - There is no "$p(w_i|c_1)$" for the multinomial case.
  - Same test phrase:

$$p(D|c_1)=p(w_1,w_2|c_1)$$
$$=p(w_1|c_1).\,p(w_2|c_1)$$
$$=\frac{3}{6}\times\frac{1}{3}=\frac{1}{6}$$

  - i.e. for MBE, each *document* is an "event", while for ME, each *word* is an "event"

Masdar INSTITUTE