CIS501 – Lecture 6.5 (replacement class)

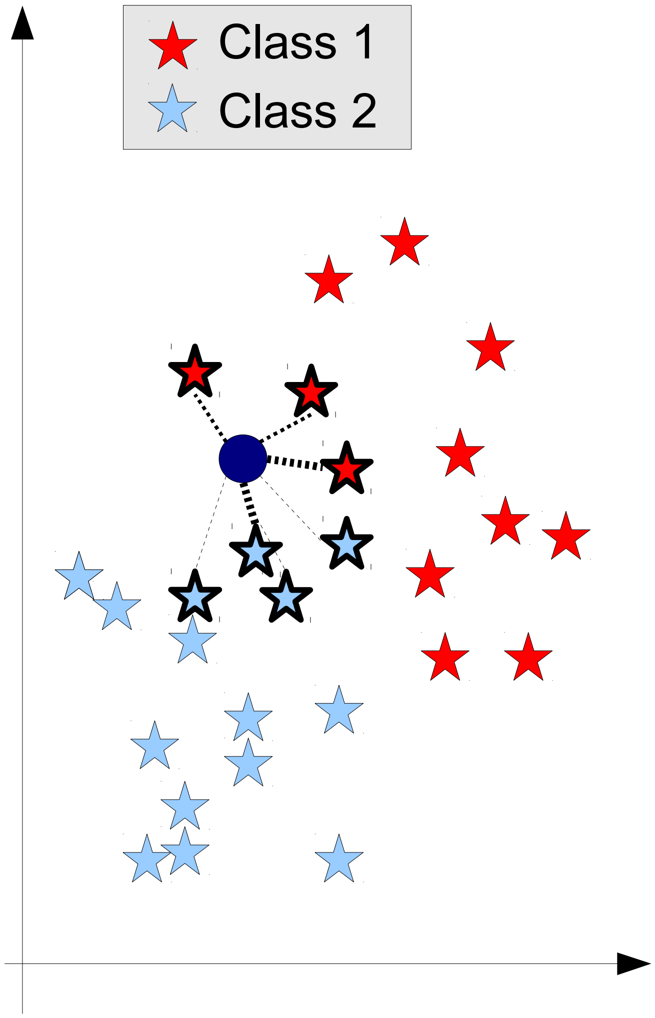
Woon Wei Lee
Fall 2013, 10-11:15am,
Sundays and Wednesdays

Masdar INSTITUTE

For today:

- Administrative stuff
  - Presentation slides
  - Lab submissions
- Kernel density estimation
- Naïve Bayes Classifier
  - Multinomial event model
- Presentations:
  - Maryam Almehrezi
  - Ya-Chen Chang

Distance weighted k-NN classifier



- Standard *k*-NN:
  - Big *k* → good noise resistance, poor resolution
  - Small *k* (the opposite)
- One trick is to emphasize closer neighbors:
  - Assign different weightings to the neighbors
  - Different weighting schemes available have been suggested:

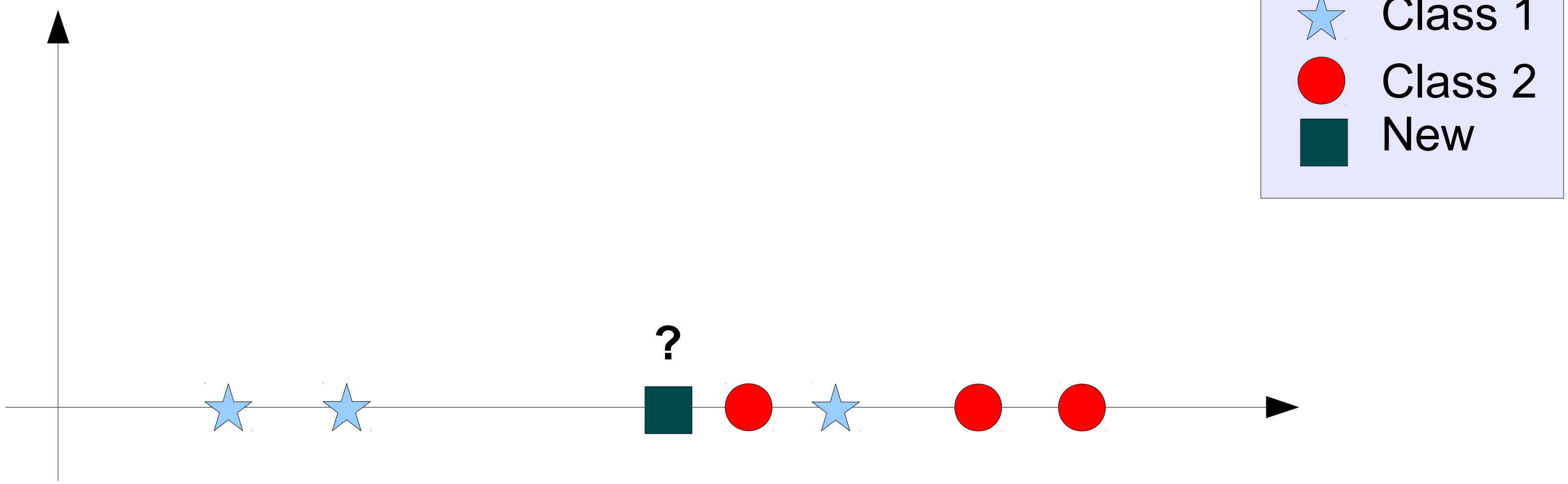
i. $$w_j = \frac{d_k - d_j}{d_k - d_1}$$

ii. $$w_j = \frac{1}{d_j}$$

iii. $$w_j = k - j + 1$$

Cont'd

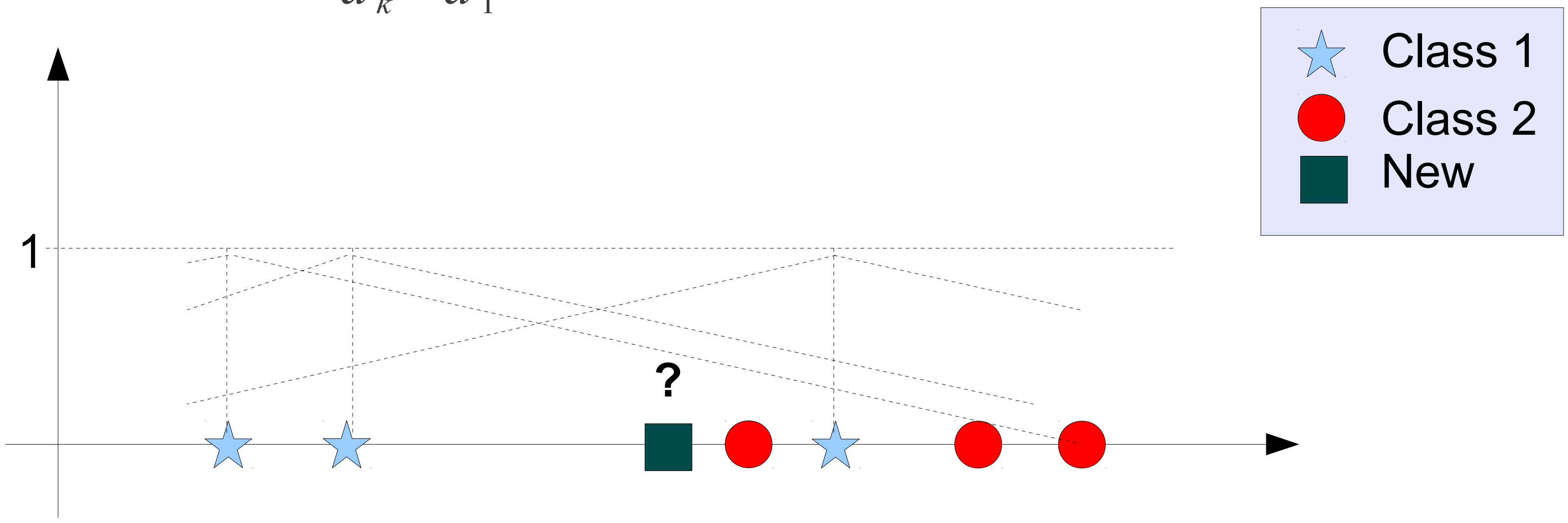
$$w_j = \frac{d_k - d_j}{d_k - d_1}$$



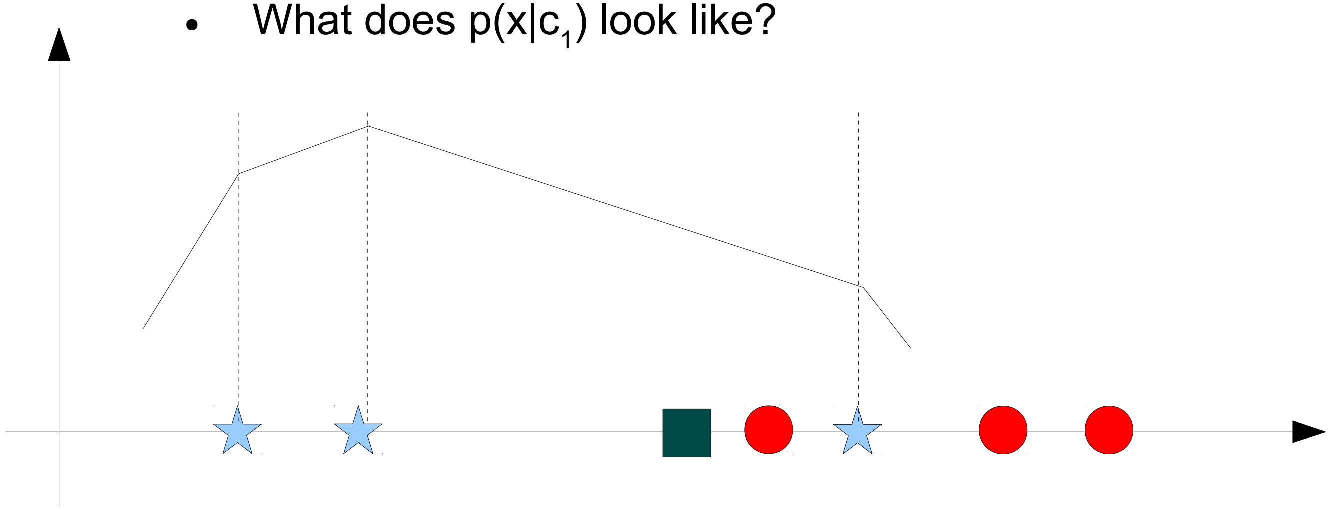
- **1-D Example**
    - Set k=6 (not normal..)

(Class 1)

$$w_j = \frac{d_k - d_j}{d_k - d_1}$$

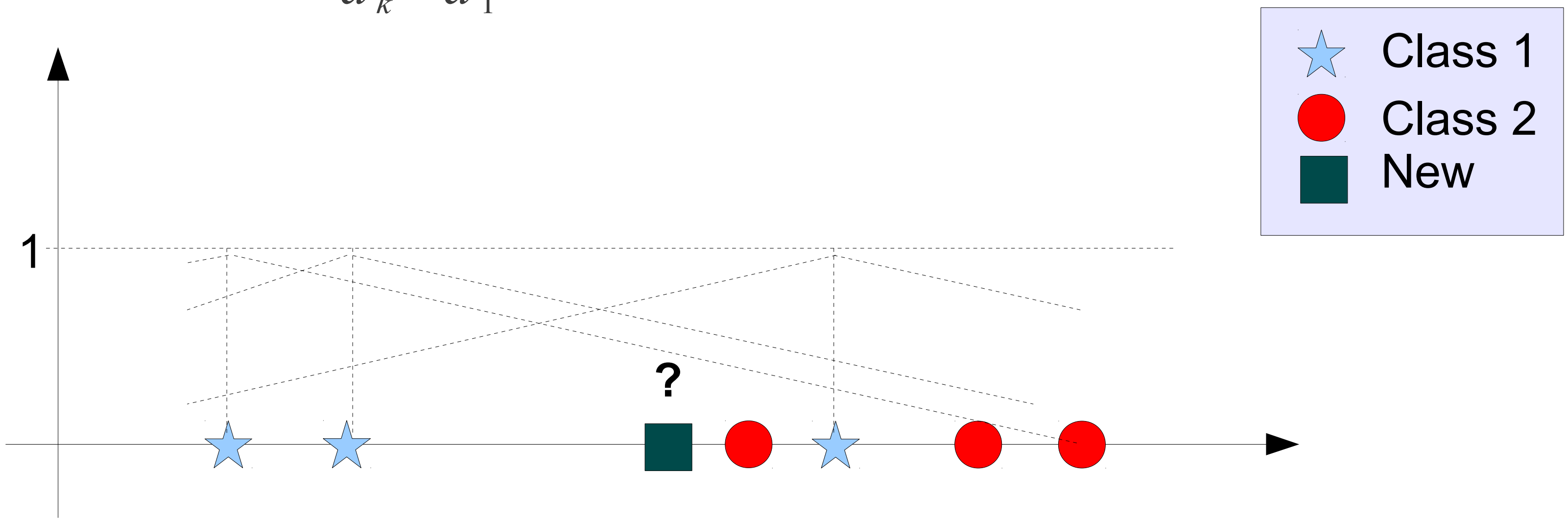


- What does $p(x|c_1)$ look like?

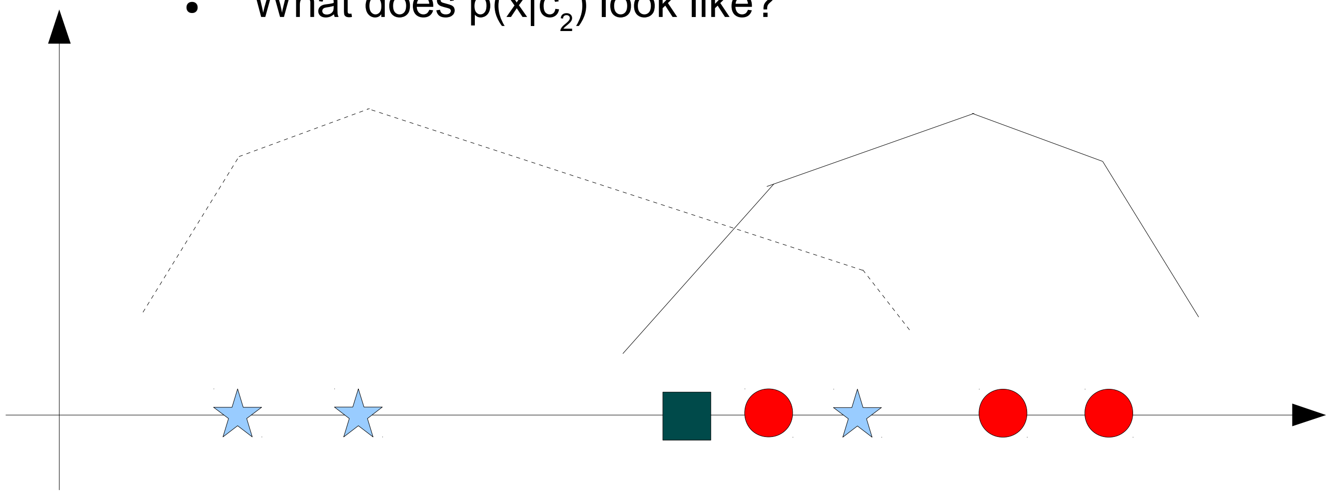


(Class 1)

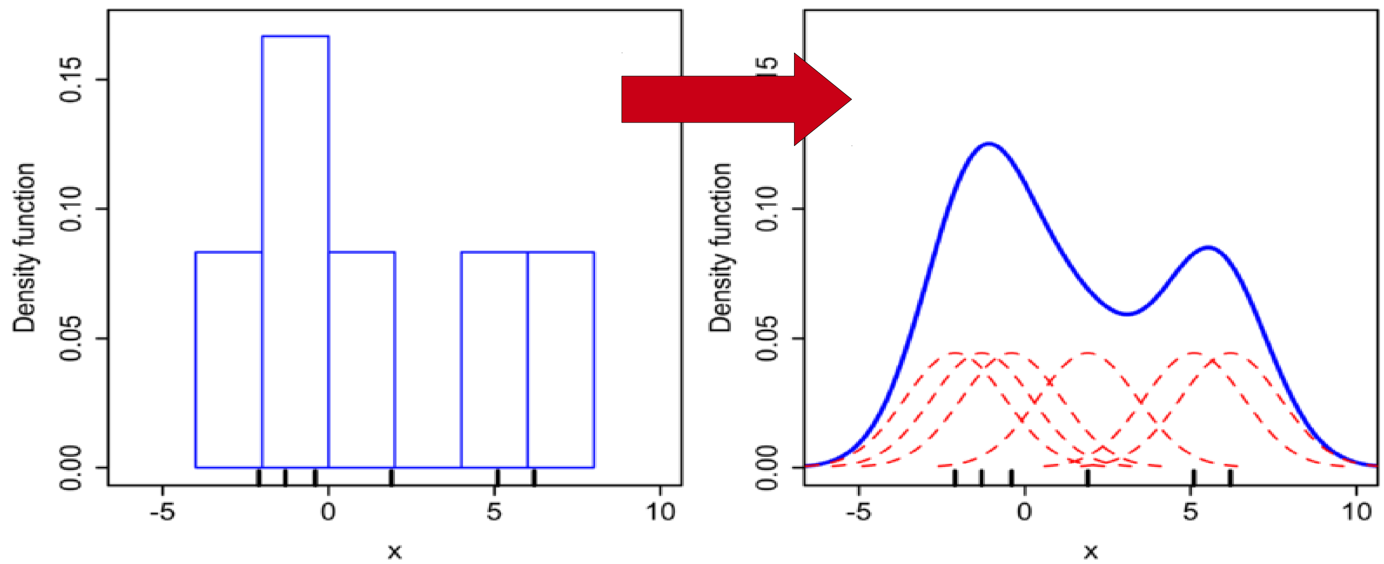
$$w_j = \frac{d_k - d_j}{d_k - d_1}$$



- What does $p(x|c_2)$ look like?



k-NN as a form of kernel density estimation

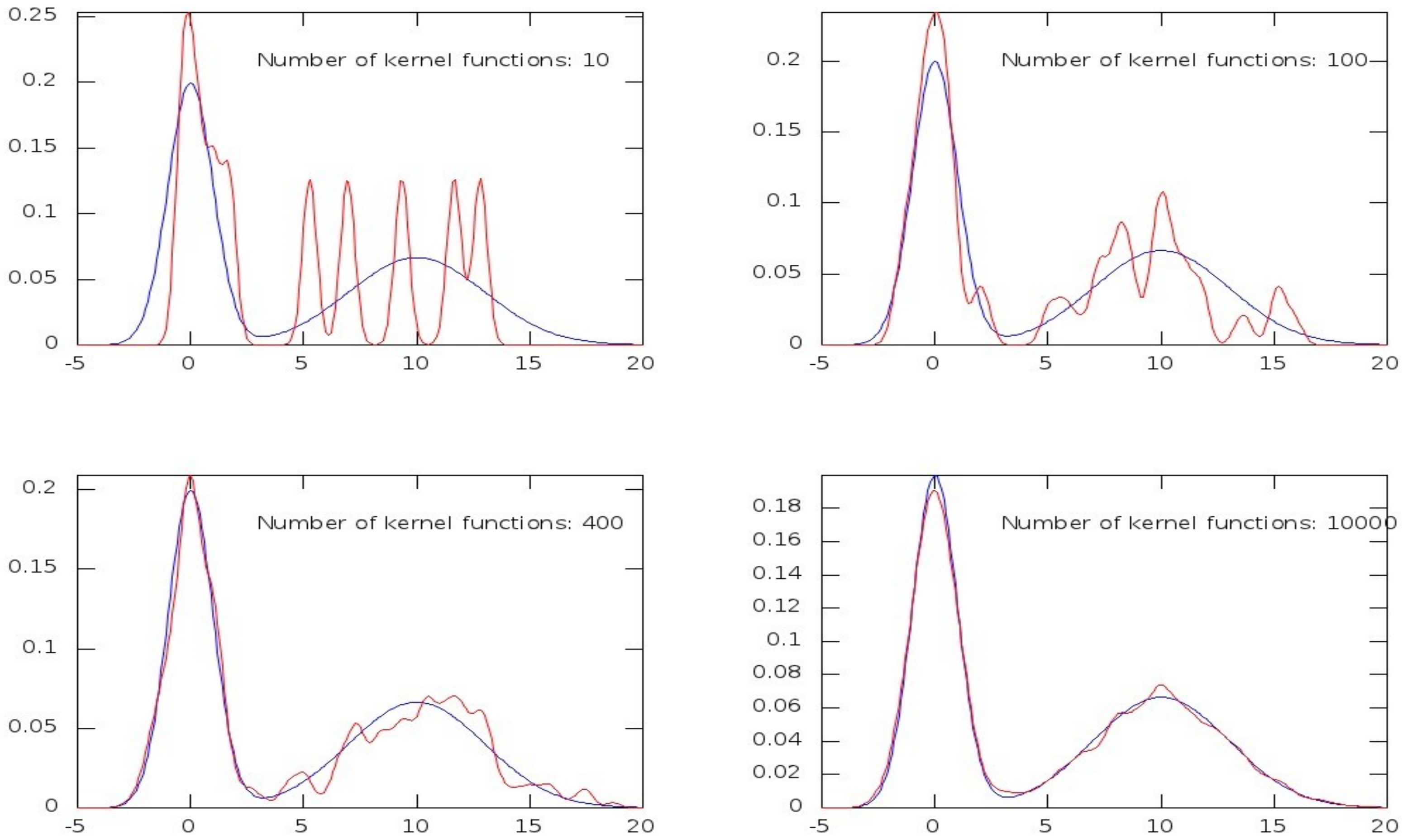


- Standard (unweighted) k-NN assumes that each of the k closest points contributes a uniform probability density to p(x|c).
- The distance weighted k-NN assumes a unimodal density (depending on weighting function).
- Related to the technique of kernel density estimation is a technique where PDF is approximated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

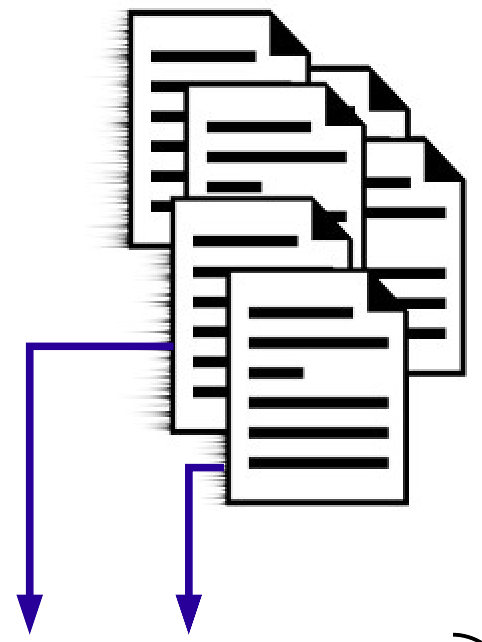


- **Example of kernel density estimation.**
  - “True density” → mixture of two gaussians, N(0,1) and N(10,3)
  - Kernels → gaussians with std of 0.1
- **Increase in number of kernel functions → greater smoothness**
- **Problem → high dimensional spaces...**

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"

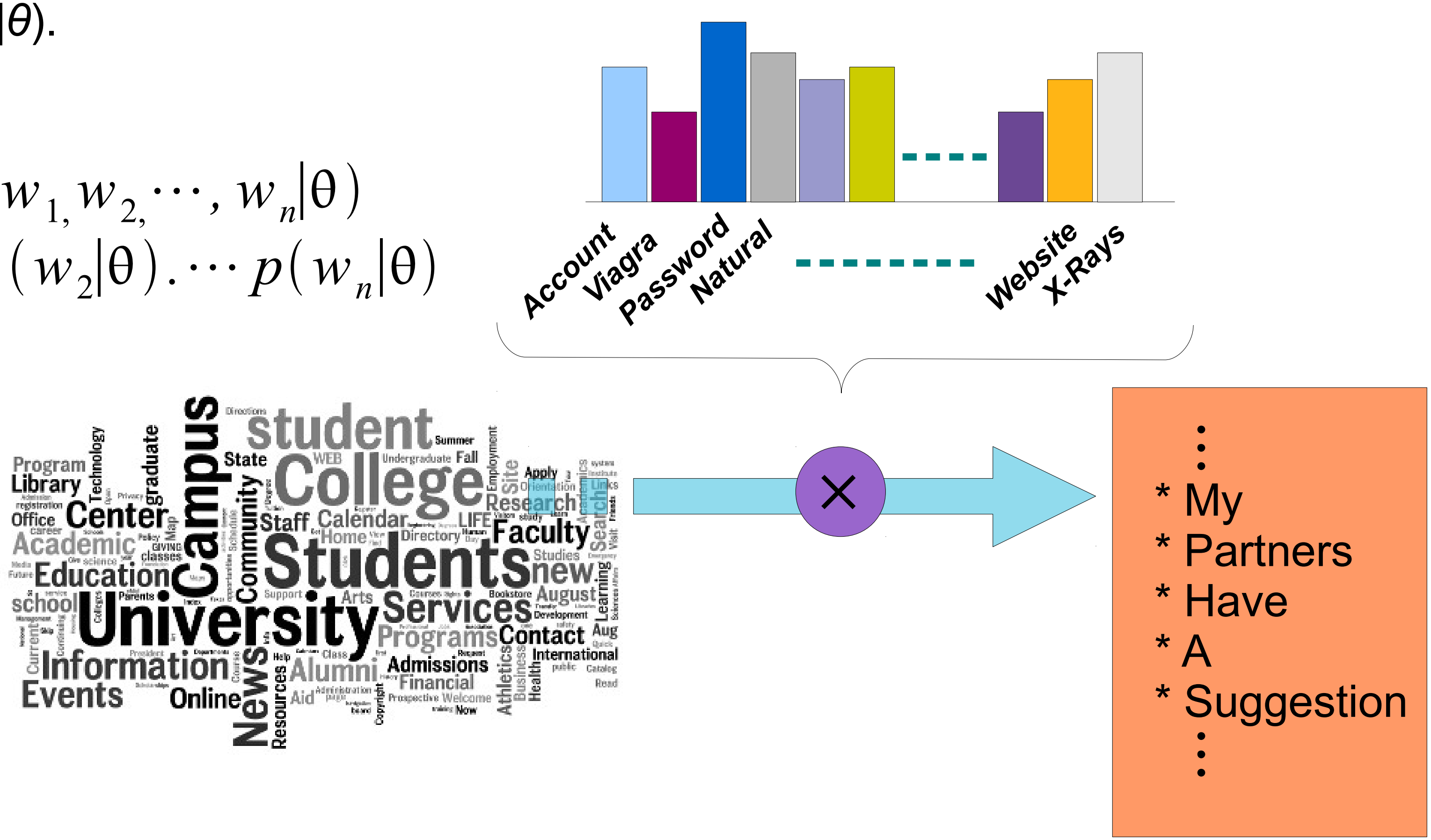


$$\begin{pmatrix} p(w_1^1|c) & p(w_1^2|c) & \\ p(w_2^1|c) & p(w_2^2|c) & \\ p(w_3^1|c) & p(w_3^2|c) & \cdots \\ \vdots & \vdots & \\ p(w_i^1|c) & p(w_i^2|c) & \end{pmatrix}$$

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|*θ*).
- 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*}
  - $p(w_i=1|c_1)=\{5/6,2/3,3/5\}$

    $\rightarrow \quad p(w_i=0|c_1)=\{1/6,1/3,2/5\}$
  - 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"