School of Computer Science
Carnegie Mellon

10-601 Introduction to Machine Learning

Final Review

Readings:

Matt Gormley
Lecture 27
December 5, 2016

Reminders

- Final Exam
  - Wed, Dec. 7th, in-class
- Final Exam Review Session
  - Tue, Dec. 6th at 5:30pm

Outline

1. Exam Logistics
2. Sample Questions
3. Overview

EXAM LOGISTICS

Final Exam

- **Exam Details**
  - In-class exam on Wed, Dec. 7th
  - 7 problems
  - Format of questions:
    - Multiple choice
    - True / False (with justification)
    - Derivations
    - Short answers
    - Interpreting figures
  - No electronic devices
  - You are allowed to **bring** one 8½ x 11 sheet of notes (front and back)

Final Exam

- **How to Prepare**
  - Attend the final recitation session: Tue, Dec. 6th at 5:30pm
  - Review prior year's exams and solutions (we will post them)
  - Review this year's homework problems

- Grading scheme drops your lowest homework grade
- If you skipped one, be sure to go back to it and carefully study all the questions

Final Exam

- **How to Prepare**
  - Attend the final recitation session:
    Tue, Dec. 6th at 5:30pm
  - Review prior year's exams and solutions
    (we will post them)
  - Review this year's homework problems
  - Flip through the "What you should know" points
    (see 'More' links on 'Schedule' page of course website)

Final Exam

- **Advice (for during the exam)**
  - Solve the easy problems first (e.g. multiple choice before derivations)
    - if a problem seems extremely complicated you're likely missing something
  - Don't leave any answer blank!
  - If you make an assumption, write it down
  - If you look at a question and don't know the answer:
    - we probably haven't told you the answer
    - but we've told you enough to work it out
    - imagine arguing for some answer and see if you like it

Final Exam

- **Exam Contents**
  - 20% of material comes from topics covered **before** the midterm exam
  - 80% of material comes from topics covered **after** the midterm exam

Topics covered **before** Midterm

- Foundations
  - Probability
  - MLE, MAP
  - Optimization
- Classifiers
  - Decision Trees
  - Naïve Bayes
  - Logistic Regression
  - Perceptron
  - SVM
- Regression
  - Linear Regression
- Important Concepts
  - Kernels
  - Regularization and Overfitting
  - Sample Complexity
  - Experimental Design

Topics covered **after** Midterm

- Supervised Learning
  - Boosting
- Unsupervised Learning
  - K-means / Lloyd's method
  - Hierarchical clustering
  - PCA
  - EM / GMMs
- Neural Networks
  - Basic architectures
  - Backpropagation
  - Why Deep Nets are hard to train
  - CNNs / RNNs
- Graphical Models
  - Bayesian Networks
  - Factor Graphs
  - HMMs / CRFs
  - Learning and Inference
- Other Learning Paradigms
  - Active Learning
  - Semi-Supervised Learning
  - ~~Reinforcement Learning~~
  - ~~Collaborative Filtering~~

SAMPLE QUESTIONS

Sample Questions

1 Topics before Midterm

(a) [2 pts.] **T or F**: Naive Bayes can only be used with MLE estimates, and not MAP estimates.

(b) [2 pts.] **T or F**: Logistic regression cannot be trained with gradient descent algorithm.

(d) [2 pts.] **T or F**: Leaving out one training data point will always change the decision boundary obtained by perceptron.

Sample Questions

1 Topics before Midterm

(e) [2 pts.] **T or F**: The function $K(\mathbf{x}, \mathbf{z}) = -2\mathbf{x}^T\mathbf{z}$ is a valid kernel function.

(h) [2 pts.] **T or F**: The VC dimension of a finite concept class H is upper bounded by $\lceil \log_2 |H| \rceil$.

Samples Questions

2 K-Means Clustering

(a) [3 pts] We are given n data points, $x_1, ..., x_n$ and asked to cluster them using K-means. If we choose the value for k to optimize the objective function how many clusters will be used (i.e. what value of k will we choose)? **No justification required.**

(i) 1 (ii) 2 (iii) n (iv) $\log(n)$

Samples Questions

2.2 Lloyd's algorithm

Circle the image which depicts the cluster center positions after 1 iteration of Lloyd's algorithm.

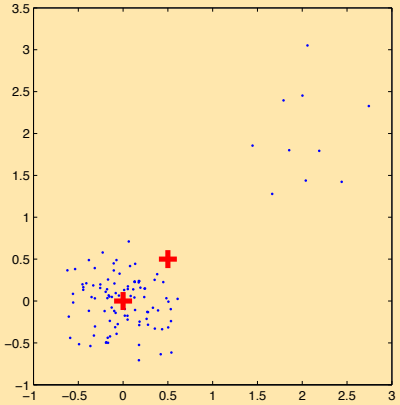
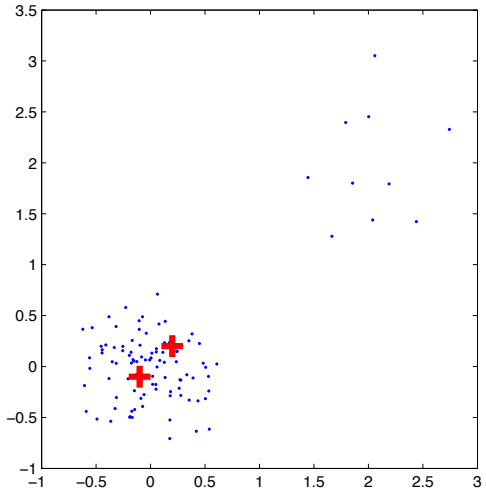
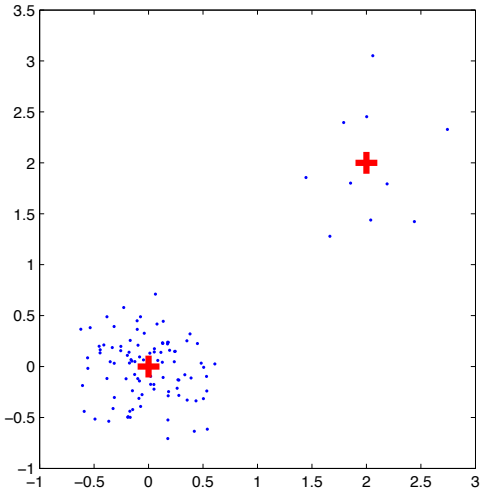
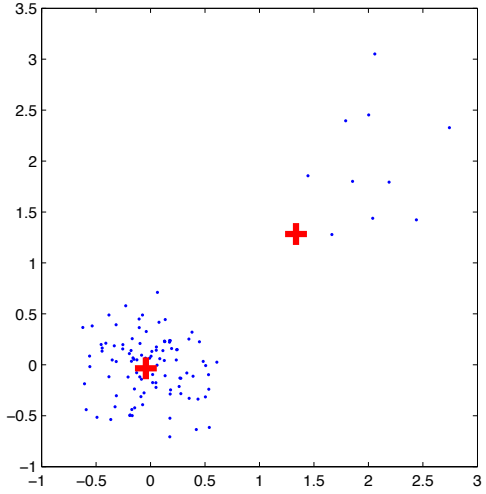
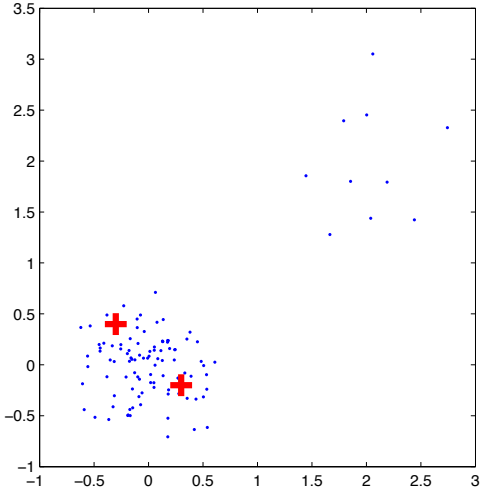


Figure 2: Initial data and cluster centers









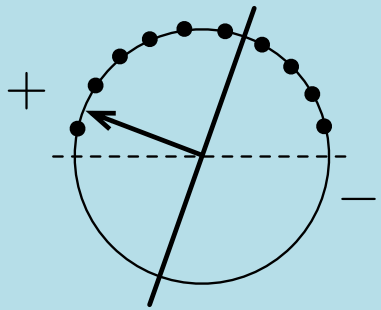
Sample Questions

3 Active Learning

3.1 Linear Separators on the Circle

In this problem you will design an active learning algorithm for finding a consistent linear separator passing through the origin when the data is on the unit circle in 2 dimensions. That is, given a dataset $S = \{x_1, \ldots, x_n\}$ with $\|x_i\| = 1$ for all $i = 1, \ldots, n$, your goal is to find a consistent classifier of the form $h(x) = \text{sign}(w^\top x)$. Assume we are in the realizable setting.

(a) [8 pts.] First, suppose that our data lies only on the *top half* of the circle (e.g., see Figure 3a). In 1–2 sentences, describe an algorithm for finding a consistent linear separator passing through the origin using $O(\log n)$ label queries. Hint: this problem is very similar to learning a consistent threshold function for data on the real line.



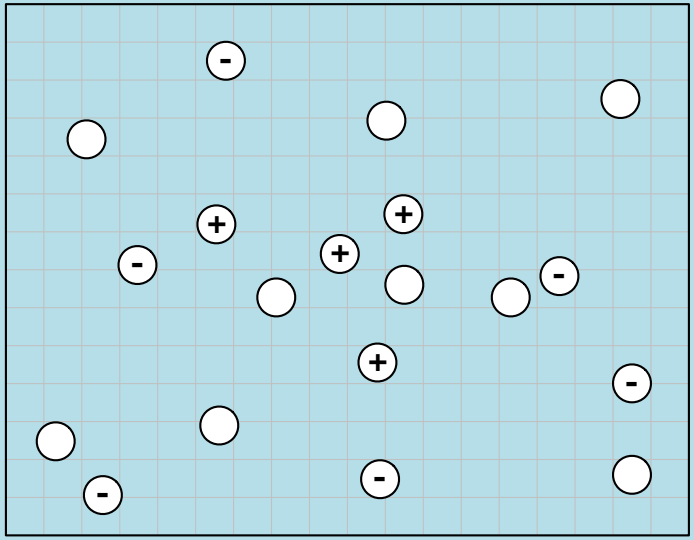
(a) Example data on the top half of the circle.

Sample Questions

3 Active Learning

(a) On Figure 5a draw the smallest and largest rectangles that correctly classify the labeled examples. In 1–2 sentences, describe the version space in terms of these two rectangles.

(b) On Figure 5a, mark each point in the region of disagreement with the letter ‘d’. For all other points, mark them with their correct label.



(a) Dataset for parts (a) and (b).

Sample Questions

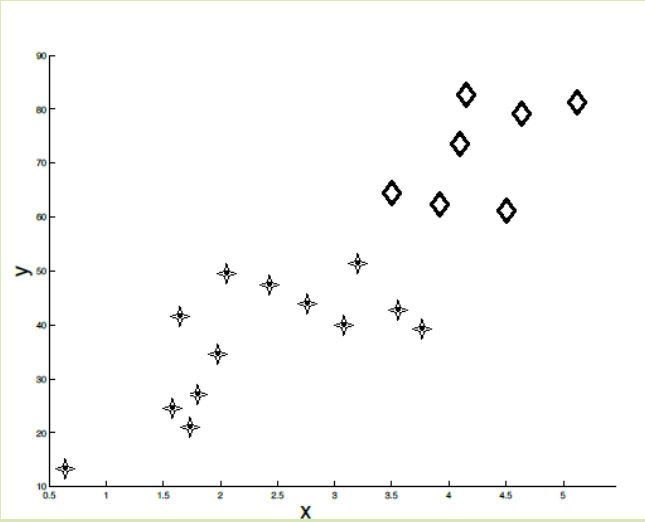
4 Principal Component Analysis [16 pts.]

(a) In the following plots, a train set of data points X belonging to two classes on $\mathbb{R}^2$ are given, where the original features are the coordinates (x, y). For each, answer the following questions:

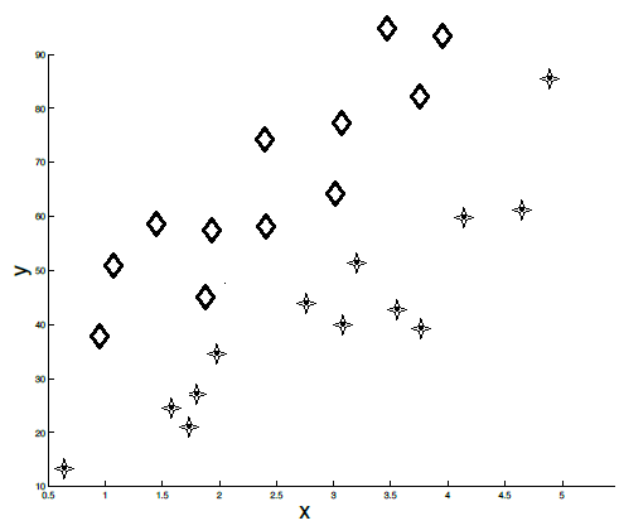
(i) [3 pt.] Draw all the principal components.

(ii) [6 pts.] Can we correctly classify this dataset by using a threshold function after projecting onto one of the principal components? If so, which principal component should we project onto? If not, explain in 1–2 sentences why it is not possible.

Dataset 1:



Dataset 2:



Sample Questions

4 Principal Component Analysis

(c) [2 pts.] Assume we apply PCA to a matrix $X \in \mathbb{R}^{n \times m}$ and obtain a set of PCA features, $Z \in \mathbb{R}^{m \times n}$. We divide this set into two, Z_1 and Z_2. The first set, Z_1, corresponds to the top principal components. The second set, Z_2, corresponds to the remaining principal components. Which is more common in the training data:

A: a point with large feature values in Z_1 and small feature values in Z_2

B: a point with large feature values in Z_2 and small feature values in Z_1

Sample Questions

(a) [2 pts.] Write the expression for the joint distribution.

5 Graphical Models [16 pts.]

We use the following Bayesian network to model the relationship between studying (S), being well-rested (R), doing well on the exam (E), and getting an A grade (A). All nodes are binary, i.e., $R, S, E, A \in \{0, 1\}$.

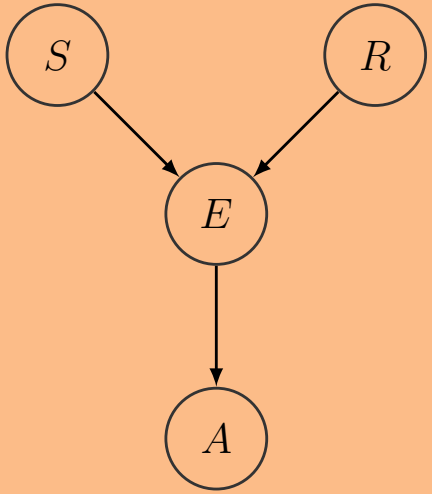


Figure 5: Directed graphical model for problem 5.

Sample Questions

(b) [2 pts.] How many parameters, i.e., entries in the CPT tables, are necessary to describe the joint distribution?

5 Graphical Models [16 pts.]

We use the following Bayesian network to model the relationship between studying (S), being well-rested (R), doing well on the exam (E), and getting an A grade (A). All nodes are binary, i.e., $R, S, E, A \in \{0, 1\}$.

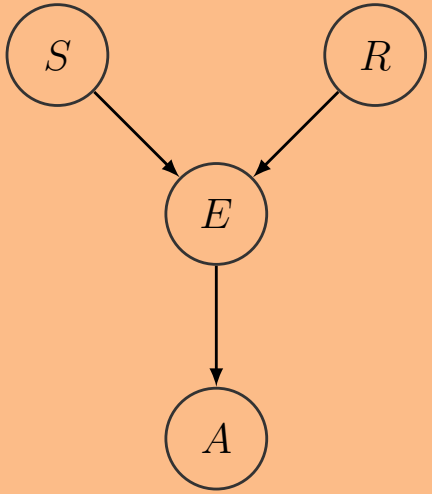


Figure 5: Directed graphical model for problem 5.

Sample Questions

(d) [2 pts.] Is S marginally independent of R? Is S conditionally independent of R given E? Answer yes or no to each questions and provide a brief explanation why.

5 Graphical Models [16 pts.]

We use the following Bayesian network to model the relationship between studying (S), being well-rested (R), doing well on the exam (E), and getting an A grade (A). All nodes are binary, i.e., $R, S, E, A \in \{0, 1\}$.

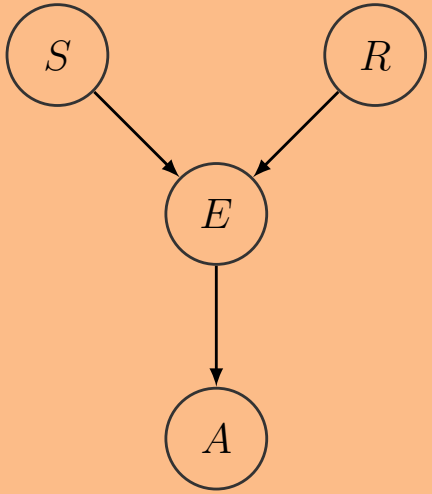


Figure 5: Directed graphical model for problem 5.

Sample Questions

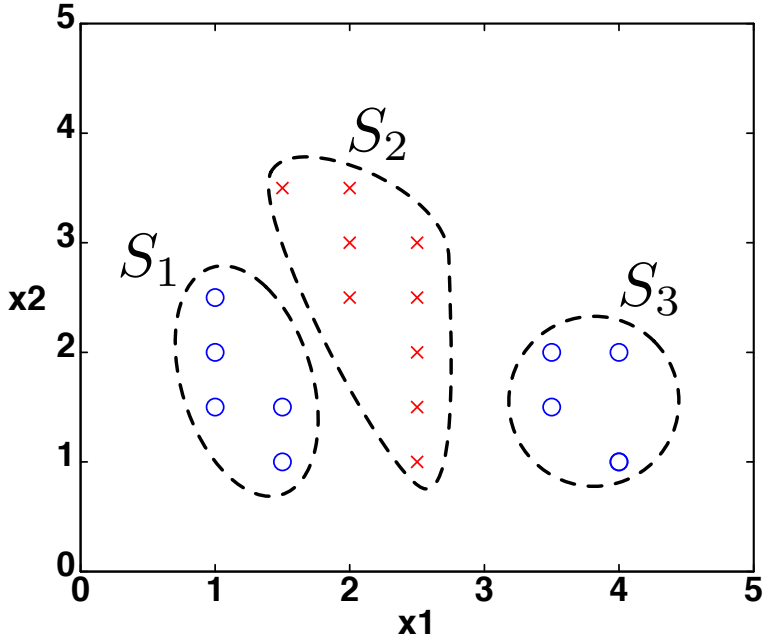
5 Graphical Models

(f) [3 pts.] Give two reasons why the graphical models formalism is convenient when compared to learning a full joint distribution.

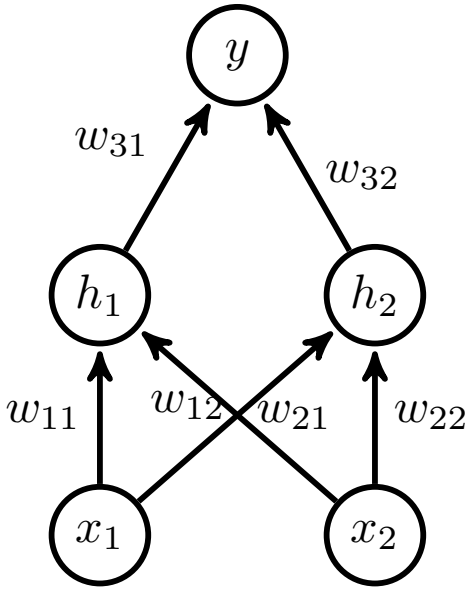
Sample Questions

Neural Networks

Can the neural network in Figure (b) correctly classify the dataset given in Figure (a)?



(a) The dataset with groups S_1, S_2, and S_3.

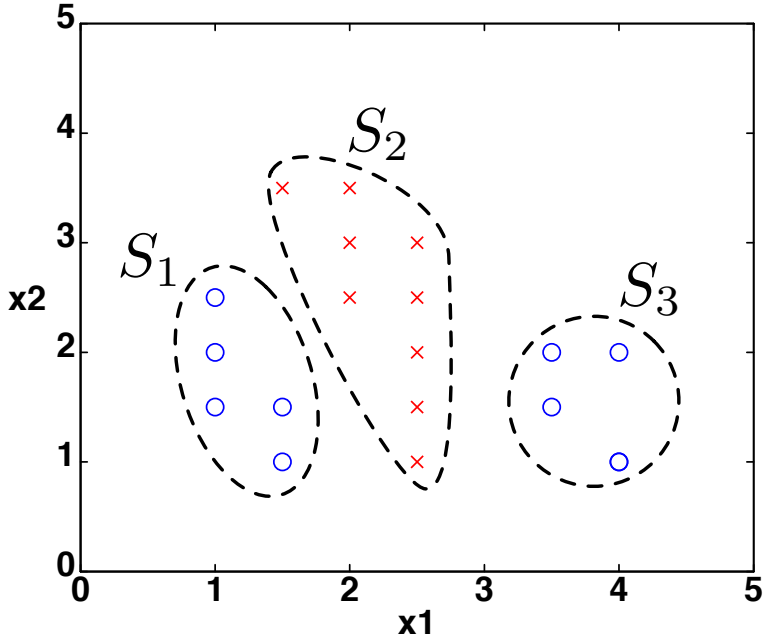


(b) The neural network architecture

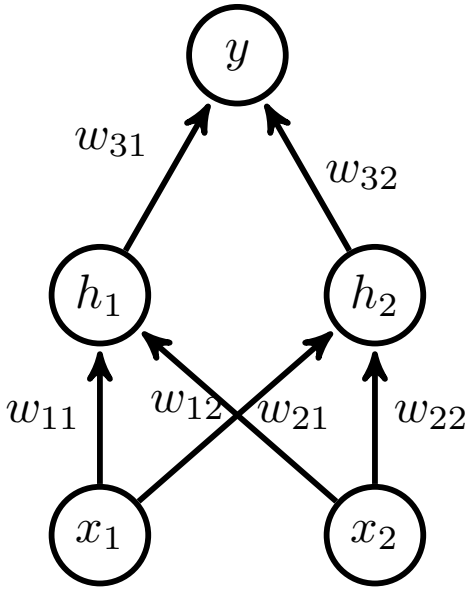
Sample Questions

Neural Networks

Can the neural network in Figure (b) correctly classify the dataset given in Figure (a)?



(a) The dataset with groups S_1, S_2, and S_3.

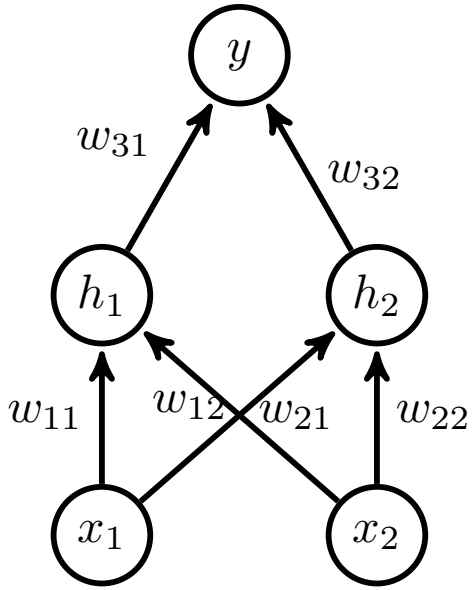


(b) The neural network architecture

Sample Questions

Neural Networks

Apply the backpropagation algorithm to obtain the partial derivative of the mean-squared error of y with the true value y^* with respect to the weight w_{22} assuming a sigmoid nonlinear activation function for the hidden layer.



(b) The neural network architecture

OVERVIEW

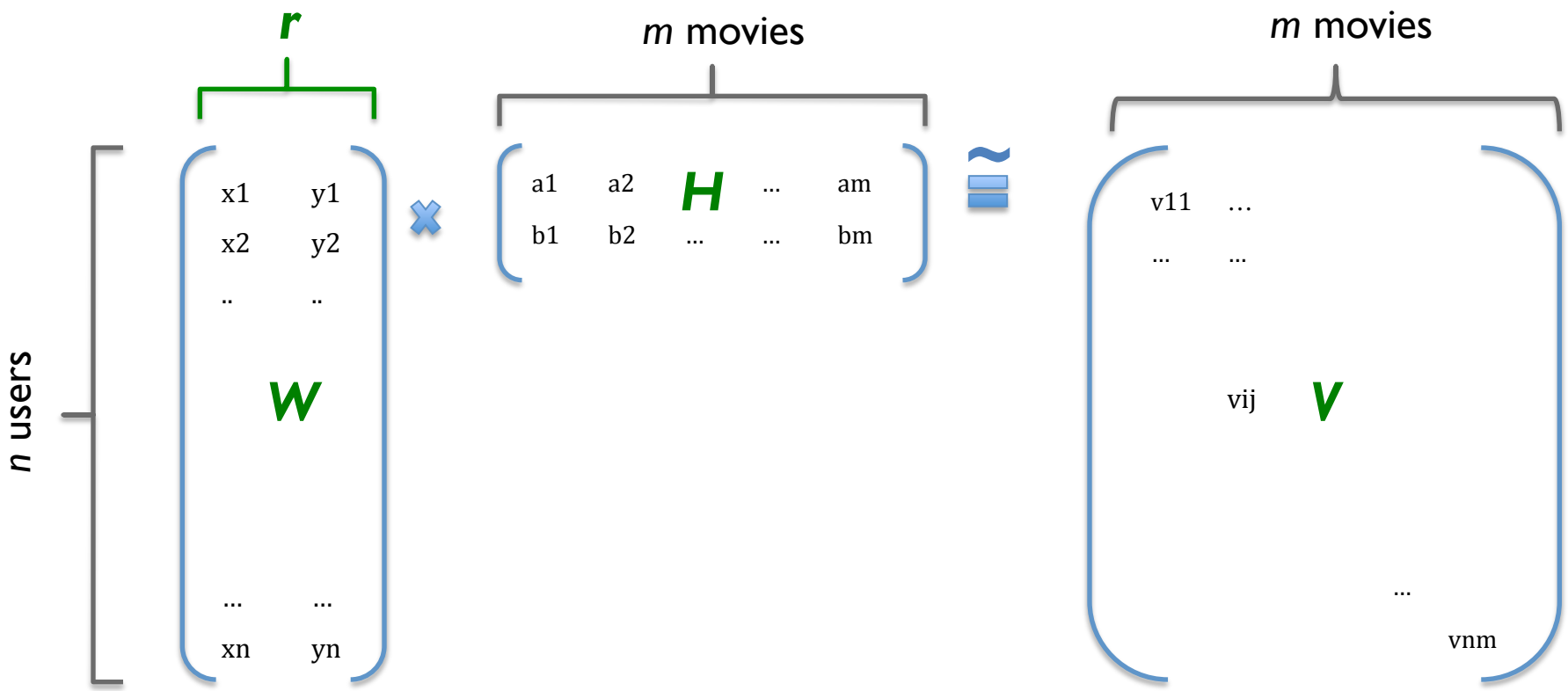
Whiteboard

- Overview #1: Learning Paradigms
- Overview #2: Recipe for ML

MATRIX MULTIPLICATION IN MACHINE LEARNING

Slides in this section from William Cohen

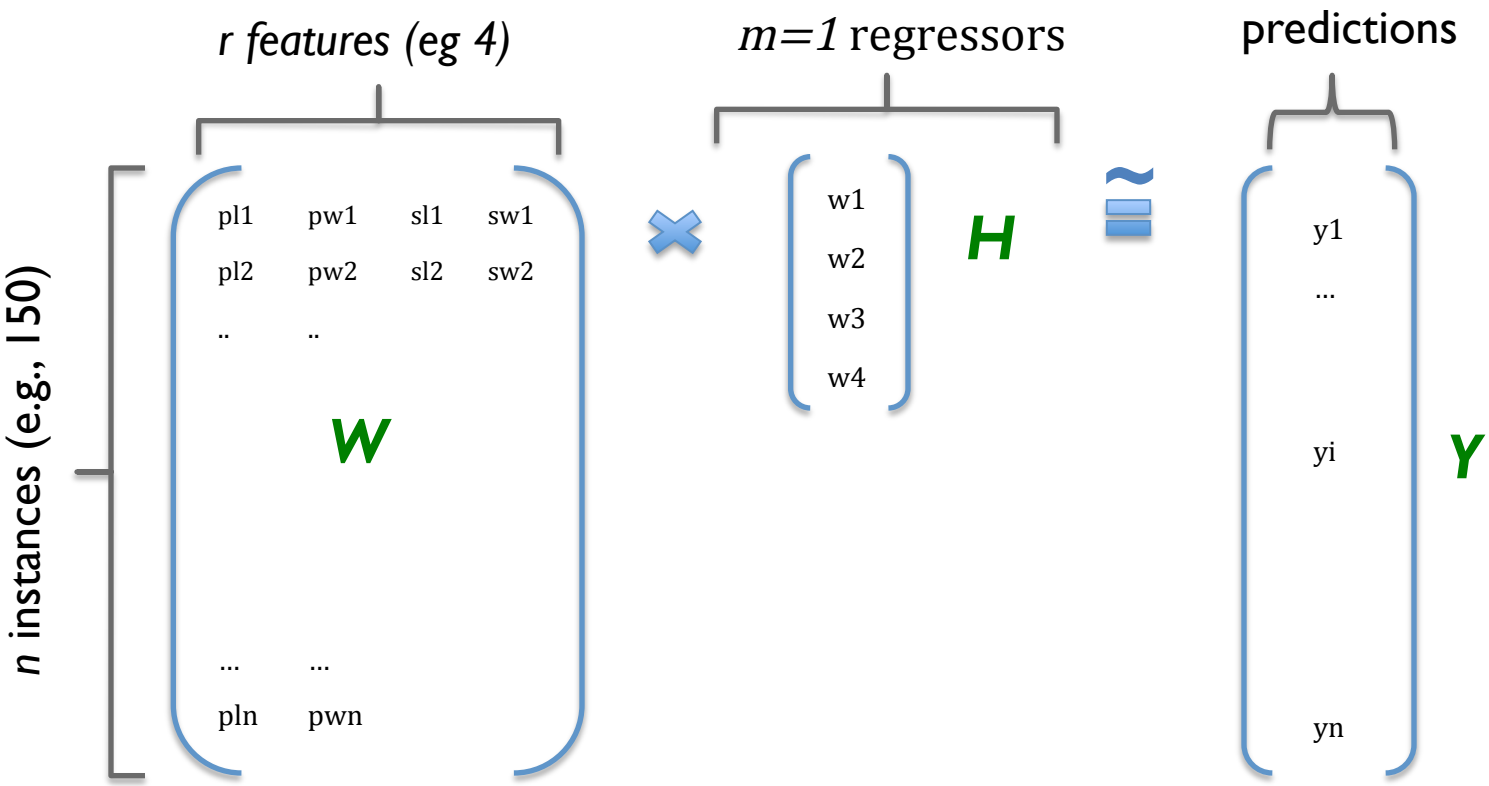
Recovering latent factors in a matrix



V[i,j] = user i's rating of movie j

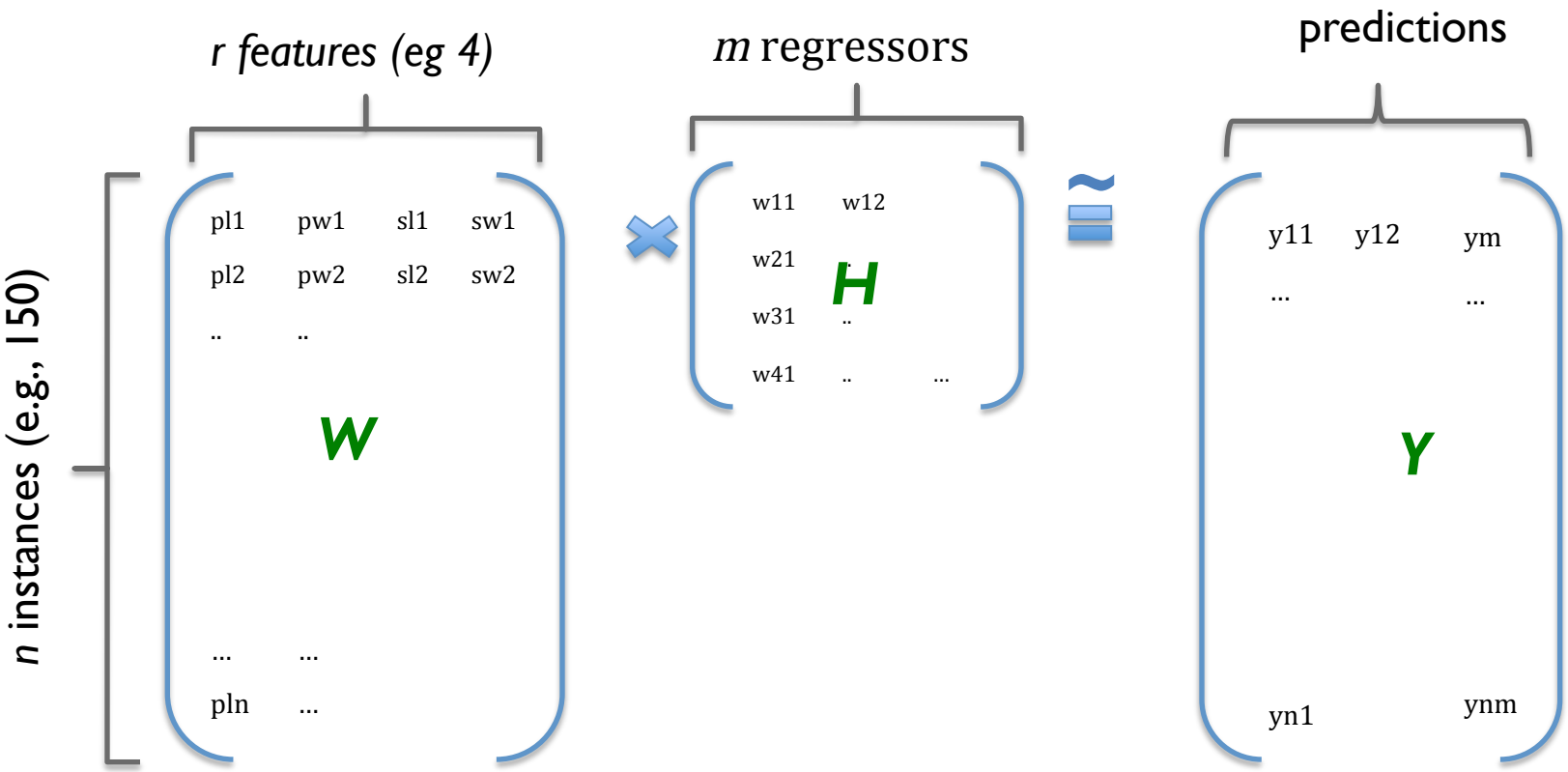
Slide from William Cohen

… is like Linear Regression ….



Y[i,1] = instance i's prediction

.. for many outputs at once….

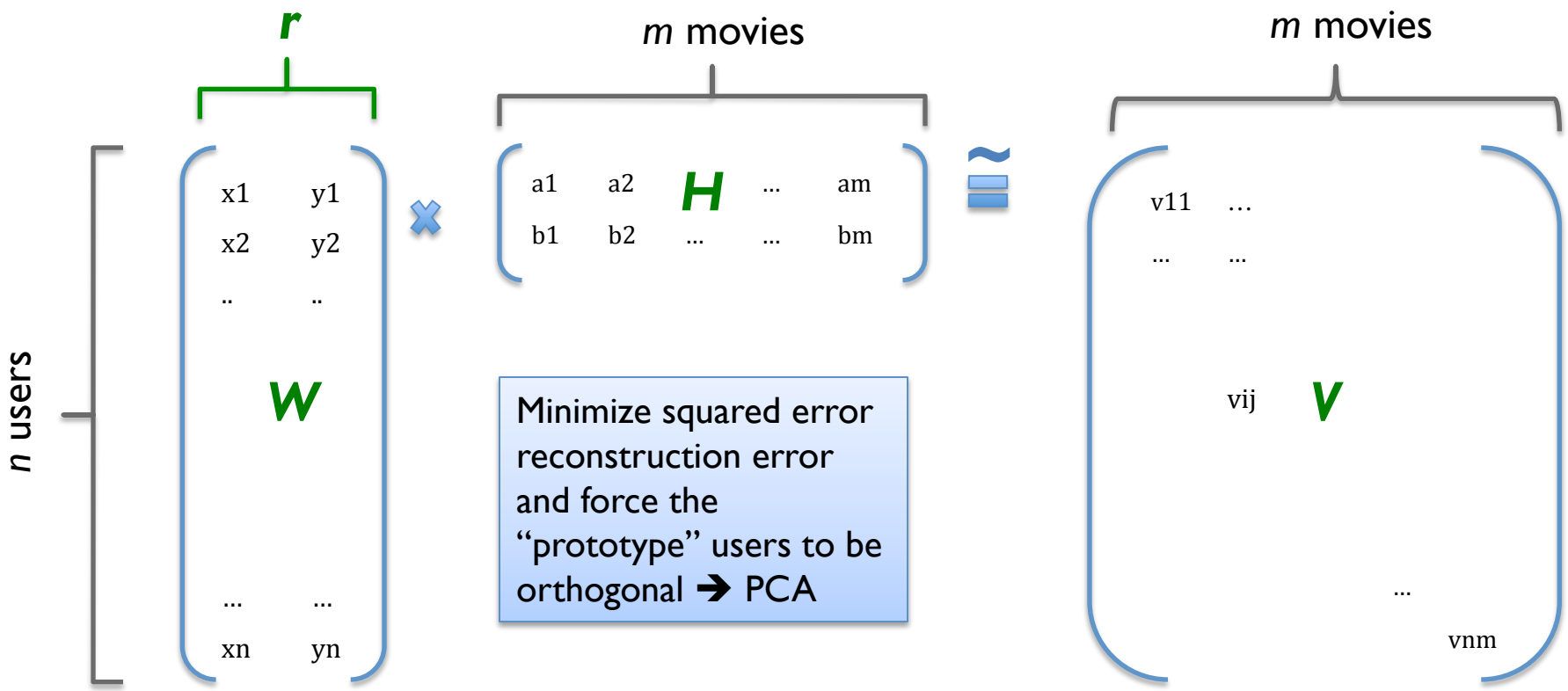


… where we also have to find the dataset!

Y[I,j] = instance i's prediction for regression task j

Slide from William Cohen

… vs PCA



V[i,j] = user i's rating of movie j

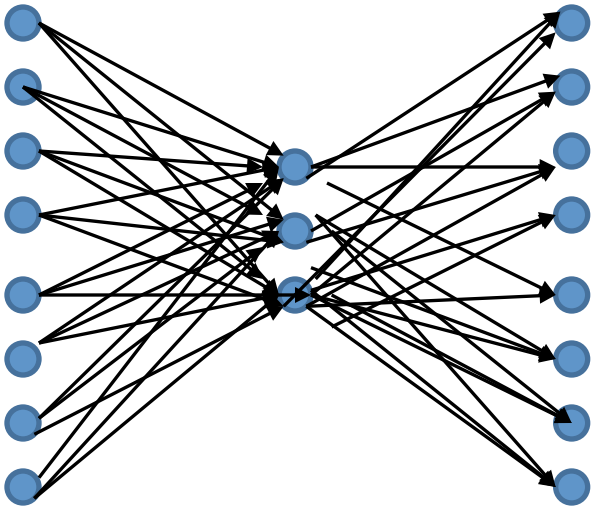
Slide from William Cohen

… vs autoencoders & nonlinear PCA

- Assume we would like to learn the following (trivial?) output function:

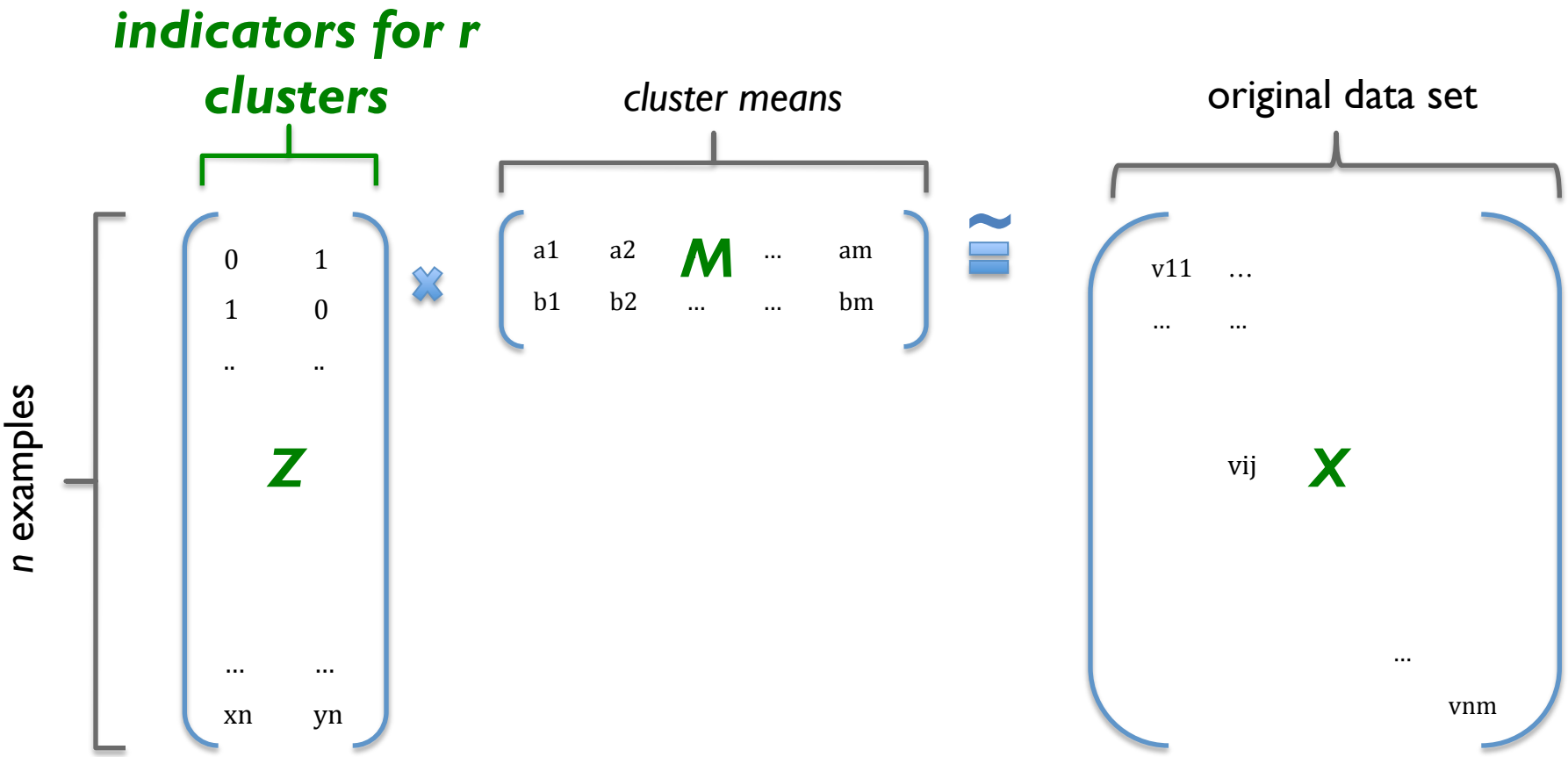
Input	Output
00000001	00000001
00000010	00000010
00000101	00000100
00001000	00001000
00010000	00010000
00100000	00100000
01000000	01000000
10000000	10000000

- Using the following network:



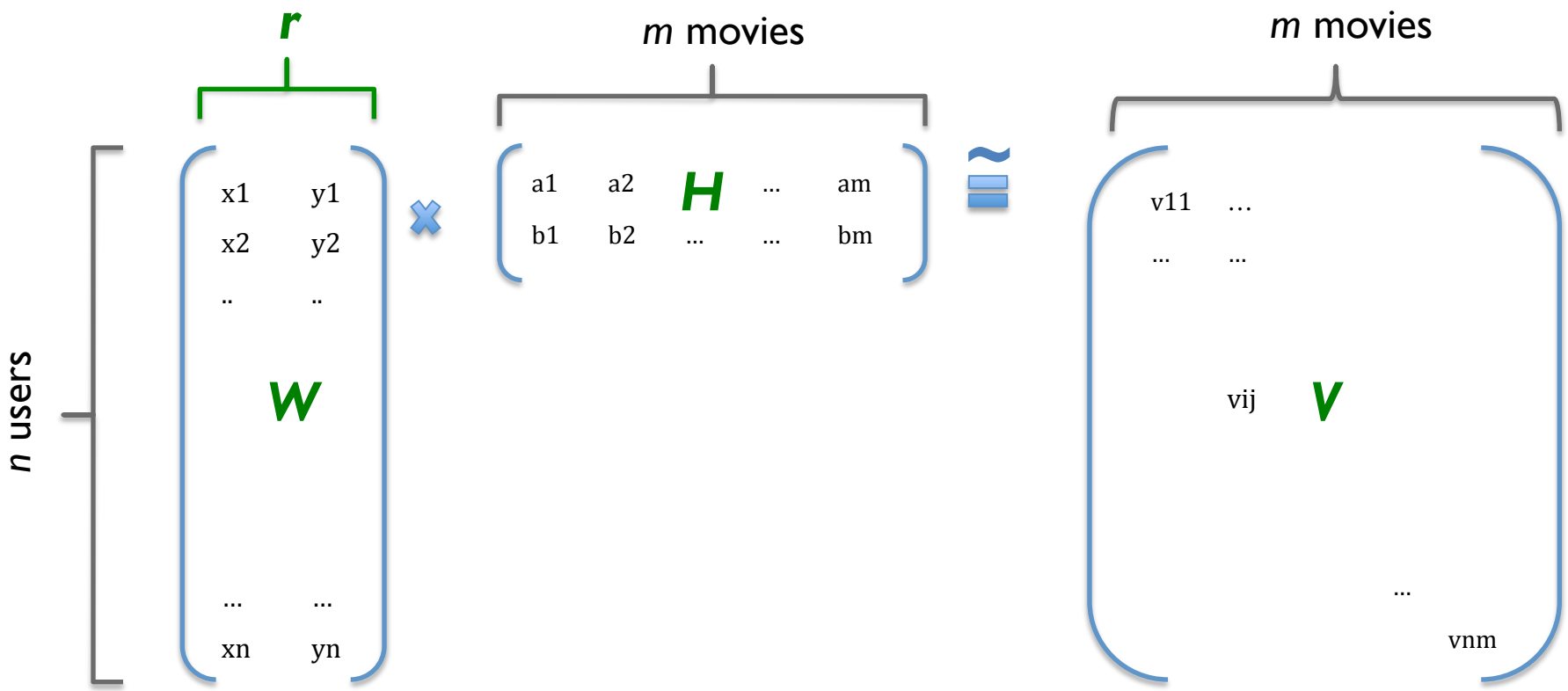
- With *linear* hidden units, how do the weights match up to W and H?

..... vs k-means



Slide from William Cohen

Recovering latent factors in a matrix



V[i,j] = user i's rating of movie j

Slide from William Cohen