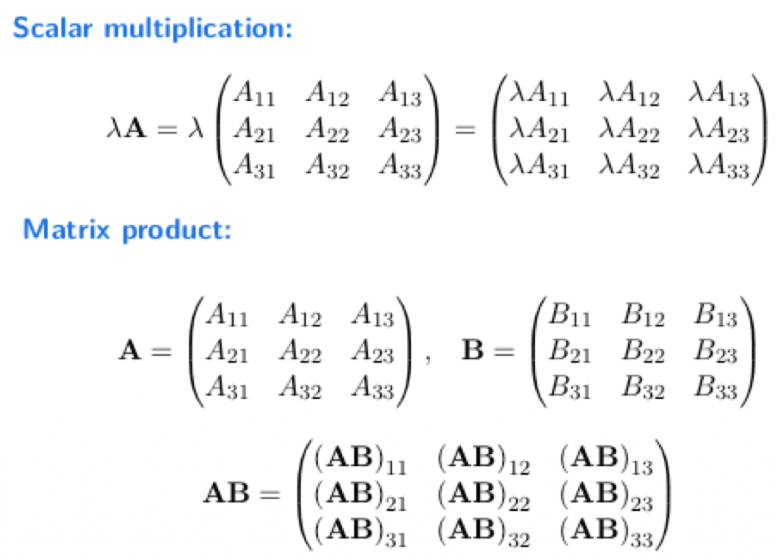
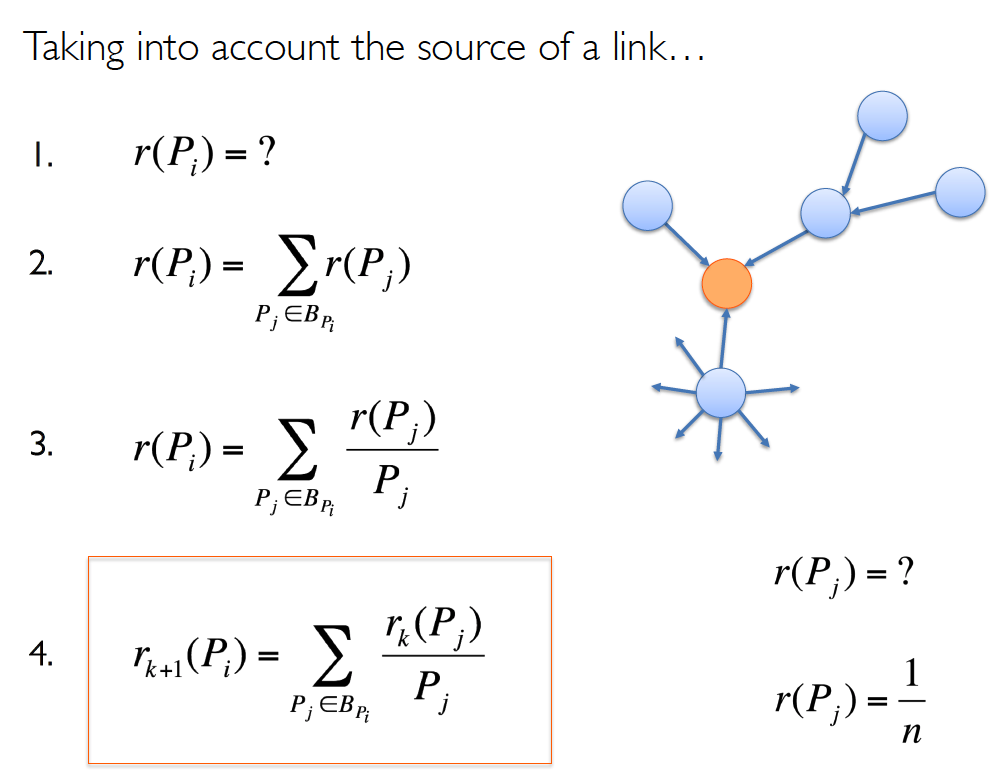
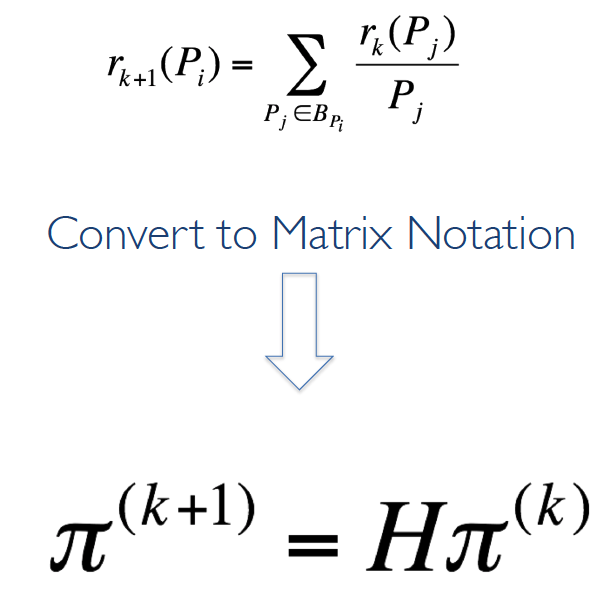
1. Nearest Neighbors
2. Designing ML experiments, training & test data
3. Supervised and unsupervised
4. Cross validation
5. PageRank
6. Linear Algebra
7. Python and data structures
8. Readings
   1. Green (Linear Algebra Appendix, pg 803 - 819) and anything from lecture/lab
   2. Page, Lawrence, et al. The PageRank citation ranking: Bringing order to the web.
   3. P. Domingos “A Few Useful Things to Know about Machine Learning”
   4. \*Data Mining, Whitten et al. (Chpts. 5, especially 5.2-5.4)
   5. Daume (Chpt. 3.1-3.3)
9. What is machine learning?
   1. Machine learning is a scientific discipline that explores the construction and study of algorithms that can **learn from data**. Such **algorithms** operate by **building a model** based on inputs and using that to make predictions or decisions, rather than following only explicitly programmed instructions.
   2. **Data mining** and **predictive analytics**. All use data to **predict** some variables as a function of other variables.
   3. A method of data analysis that **automates analytical model building**. The idea that systems can **learn from data**, identify **patterns** and make **decisions** with minimal human intervention
   4. Every ML algorithm has three primary components
      1. **Representation** (i.e., model)
      2. **Evaluation** (i.e., an objective function)
      3. **Optimization** (e.g., search)
10. What is econometrics?
    1. Use statistical methods for prediction, inference, causal modeling of economic relationships.
11. Page Rank
    1. **Degree**: number of direct ties
       1. Measure of overall activity or extend of involvement
       2. **High** degree positions are **influential**, but also may be subject to a great deal of influence from others.
    2. Eigenvector Centrality: values of the first eigenvector of the graph adjacency matrix
       1. Measure of “**coreness”** in a network
       2. 
       3. Suppose that page Pj has lj links. If one of those links is to page Pi, then Pj will pass on 1/lj of its importance to Pi.
          1. The importance ranking of Pi is the sum of all the contributions made by pages linking to it
       4. 
       5. 
12. Statistical model
    1. Good model
       1. All models has limitation, by the validity of the assumptions on which they ride.
       2. **R2  increase means good fit, but doesn’t mean a good model**
    2. Regression
       1. Statistics/econometrics
          1. Explaining **relationships**
          2. Understanding **causality (因果关系)**
          3. Does more education lead to higher wages
       2. ML/computer science
          1. **Predictions**
          2. Looking at more generalizable **patterns**
          3. How much would I earn if I left the university?
13. Supervised vs. Unsupervised
    1. Supervised
       1. We know the right answer for some values
       2. Goal is typically to model the relationship between input variables and know output variables
       3. Examples
          1. Classifying hand-written numbers from images
          2. Identifying presidential candidates based on their speeches
          3. Determining whether a student will graduate from UW
          4. Disease classification
          5. Credit scoring
    2. Unsupervised learning
       1. We don’t know the right answer, the right groupings or ground truth
       2. Goal is typically to discover underlying structure in the data
       3. Often more explanatory than supervised learning, which is more directed
14. **Overfitting**: If we have too many features, our model may fit some data really well but not handle other new examples very well at all.
15. Key issues in (Supervised) Machine Learning
    1. Don’t forget the theory
    2. Generalization and overfitting
    3. Feature Engineering
    4. More Data Matters
    5. Ensembles work
    6. Interpretability is important
16. Designing of ML experiments: Key Concepts
    1. Generalization and Overfitting
       1. Adding more features generally improves the “fit” of your model to your data
    2. Train, test, held-out data
       1. Test set helps avoid overfitting
       2. Validation
          1. Splitting into training + testing is often not enough
             1. **Hyperparameters** must be chosen
             2. Model selection, feature selection, etc.
             3. Each time you look at the test set, you introduce bias (in yourself!)
          2. Validation data
             1. A third split of the data
             2. Aka “development” data
             3. Used to fit model
             4. k-fold cross-validation

Randomly partition data into k equal size subsamples

Use each of k folds as validation set once

Average performance across k test runs

* + - * 1. Stratified(分层) cross-validation

Select folds so that the mean response value is approximately equal in all the folds, or so that some other parameter is balanced across folds

* 1. Hyperparameters
  2. Accuracy, ROC, AUC, F-scores
  3. Lift curves
  4. Baselines
  5. Error Analysis
  6. Ablative Analysis