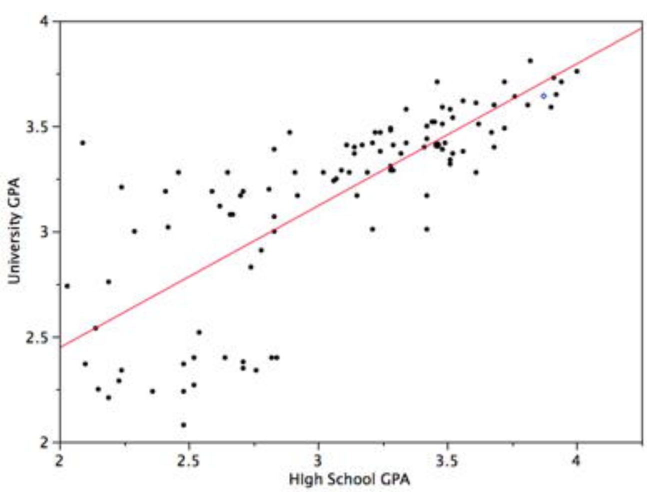
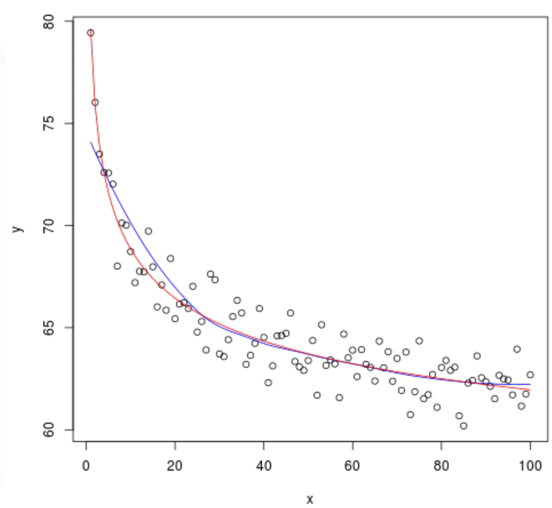
CSE 40647/60647 Data Science (Spring 2018)
Lecture 5: Data Reduction

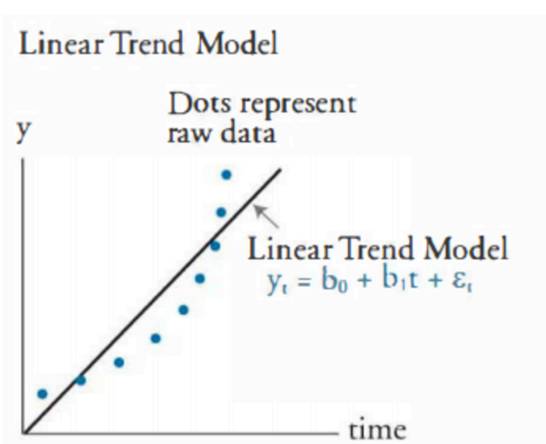
Goals:

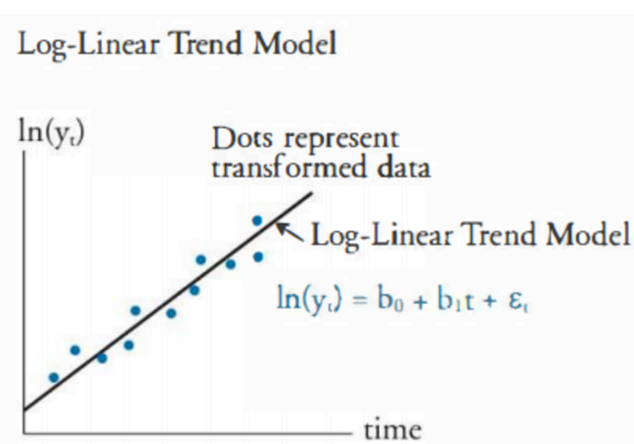
- Describe **numerosity reduction** (reducing #instances)
  - Parametric methods: Fit some model and estimate model parameters
    - Regression: Describe linear/non-linear regression models
  - Nonparametric methods
    - Histograms
    - Clustering
    - Sampling: Describe stratified sampling
- Describe **dimensionality reduction** (reducing #features)
  - Feature selection
    - Heuristic search
  - Feature extraction
    - Principal component analysis (PCA)
    - Singular Value Decomposition (SVD)

Part I: Regression models:

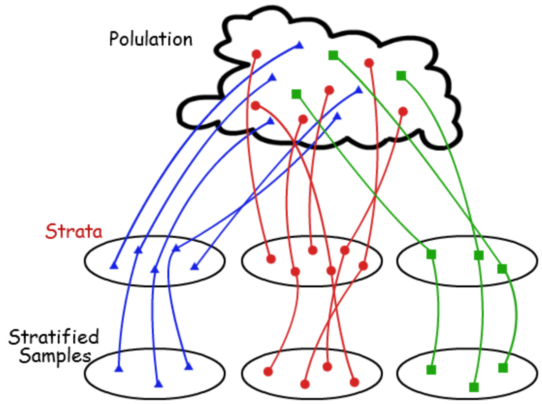




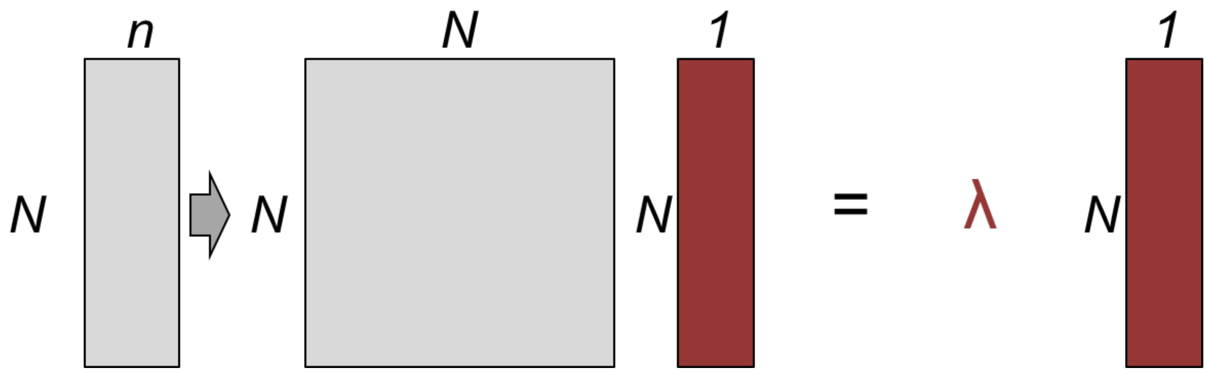


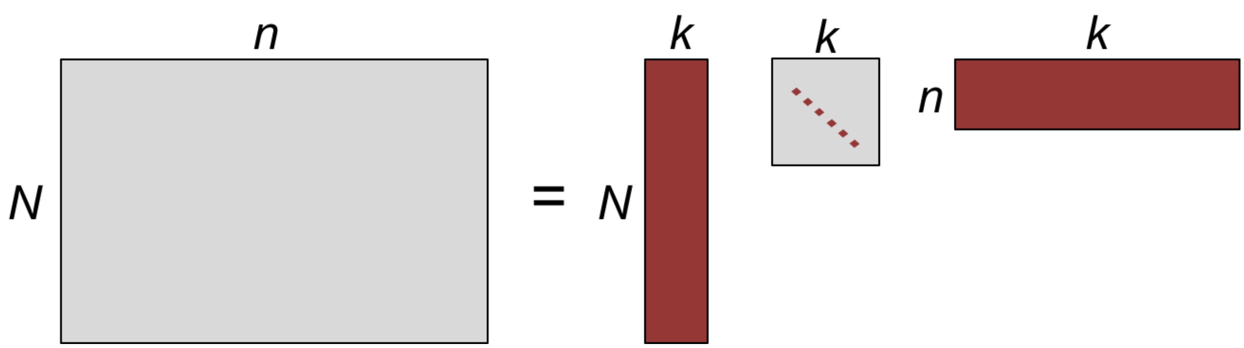


Part II: Stratified sampling



Part III: Principal Component Analysis (PCA) vs Singular Value Decomposition (SVD)





*** SVD for botnet account detection:**

Jiang, M., Cui, P., Beutel, A., Faloutsos, C. and Yang, S., 2014, May. Inferring strange behavior from connectivity pattern in social networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 126-138). Springer, Cham.

Name: **NetID:**

Please write down whatever question you have about this course:

Eigenvector computation:

$$
\begin{aligned}
\mathbf{Ax} = \lambda\mathbf{x} \quad &\Leftrightarrow \quad \mathbf{Ax} - \lambda\mathbf{x} = \mathbf{0} \\
&\Leftrightarrow \quad \mathbf{Ax} - \lambda\mathbf{Ix} = \mathbf{0} \\
&\Leftrightarrow \quad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.
\end{aligned}
$$

The equation $\mathbf{Ax} = \lambda\mathbf{x}$ has nonzero solutions for the vector x if and only if the matrix $\mathbf{A} - \lambda\mathbf{I}$ has zero determinant.

Example: Find the eigenvalues of the matrix $\mathbf{A} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix}$.

The eigenvalues are those λ for which $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. Now

$$
\begin{aligned}
\det(\mathbf{A} - \lambda\mathbf{I}) &= \det\left(\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} - \lambda\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \\
&= \det\left(\begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) \\
&= \begin{vmatrix} 2-\lambda & 2 \\ 5 & -1-\lambda \end{vmatrix} \\
&= (2-\lambda)(-1-\lambda) - 10 \\
&= \lambda^2 - \lambda - 12.
\end{aligned}
$$

The eigenvalues of $\mathbf{A}$ are the solutions of the quadratic equation $\lambda^2 - \lambda - 12 = 0$, namely $\lambda_1 = -3$ and $\lambda_2 = 4$.

First, we work with $\lambda = -3$. The equation $\mathbf{Ax} = \lambda\mathbf{x}$ becomes $\mathbf{Ax} = -3\mathbf{x}$. Writing

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and using the matrix $\mathbf{A}$ from above, we have

$$\mathbf{Ax} = \begin{bmatrix} 2 & 2 \\ 5 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix},$$

while

$$-3\mathbf{x} = \begin{bmatrix} -3x_1 \\ -3x_2 \end{bmatrix}.$$

Setting these equal, we get

$$
\begin{aligned}
\begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix} = \begin{bmatrix} -3x_1 \\ -3x_2 \end{bmatrix} \quad &\Rightarrow \quad 2x_1 + 2x_2 = -3x_1 \qquad \text{and} \qquad 5x_1 - x_2 = -3x_2 \\
&\Rightarrow \quad 5x_1 = -2x_2 \\
&\Rightarrow \quad x_1 = -\frac{2}{5}x_2.
\end{aligned}
$$

$$\mathbf{u_1} = \begin{bmatrix} 2 \\ -5 \end{bmatrix}$$

Similarly, we can find eigenvectors associated with the eigenvalue $\lambda = 4$ by solving $\mathbf{Ax} = 4\mathbf{x}$:

$$\begin{bmatrix} 2x_1 + 2x_2 \\ 5x_1 - x_2 \end{bmatrix} = \begin{bmatrix} 4x_1 \\ 4x_2 \end{bmatrix} \begin{array}{l} \Rightarrow \quad 2x_1 + 2x_2 = 4x_1 \qquad \text{and} \qquad 5x_1 - x_2 = 4x_2 \\ \Rightarrow \quad x_1 = x_2. \end{array}$$

Hence the set of eigenvectors associated with $\lambda = 4$ is spanned by

$$\mathbf{u_2} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Example: Find the eigenvalues and associated eigenvectors of the matrix

$$\mathbf{A} = \begin{bmatrix} 7 & 0 & -3 \\ -9 & -2 & 3 \\ 18 & 0 & -8 \end{bmatrix}.$$

First we compute $\det(\mathbf{A} - \lambda\mathbf{I})$ via a cofactor expansion along the second column:

$$\begin{aligned} \begin{vmatrix} 7-\lambda & 0 & -3 \\ -9 & -2-\lambda & 3 \\ 18 & 0 & -8-\lambda \end{vmatrix} &= (-2-\lambda)(-1)^4 \begin{vmatrix} 7-\lambda & -3 \\ 18 & -8-\lambda \end{vmatrix} \\ &= -(2+\lambda)[(7-\lambda)(-8-\lambda)+54] \\ &= -(\lambda+2)(\lambda^2+\lambda-2) \\ &= -(\lambda+2)^2(\lambda-1). \end{aligned}$$

Thus $\mathbf{A}$ has two distinct eigenvalues, $\lambda_1 = -2$ and $\lambda_3 = 1$. (Note that we might say $\lambda_2 = -2$, since, as a root, -2 has multiplicity two. This is why we labelled the eigenvalue 1 as λ_3.)