Stat 306:
Finding Relationships in Data.

Lecture 21

6.2 Logistic regression (Part 3) + Principal Component Analysis

Maximum Likelihood Estimation

Steps to get Maximum Likelihood Estimates:

1. Define the Likelihood as function of the parameters given the data.
2. Define the log-Likelihood.
3. Maximize the log-Likelihood or minimize the negative log-Likelihood.

For logistic regression, we have:

logistic negative log-likelihood

$$-\log L(\boldsymbol{\beta}; \text{data}) = -\sum_{i=1}^{n}(\mathbf{x}_i^T\boldsymbol{\beta})y_i + \sum_{i=1}^{n}\log[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}]$$

Unfortunately…

There is no closed form solution, but statistical software obtain $\hat{\beta}_0, \hat{\beta}_1$ with an iterative method.

Maximum Likelihood Estimation

Steps to get standard error for the Maximum Likelihood Estimates (MLE):

1. Take the second derivative of the negative log-Likelihood function. This is the Hessian Matrix. The negative of the Hessian is the Fisher information matrix.

2. Evaluate the Fisher Information matrix at the MLE. This is known as the "observed Fisher Information matrix".

3. Take the inverse of the "observed Fisher Information matrix". This is your estimate for the Variance-Covariance matrix of your parameter. The diagonal elements are the estimated Variances.

4. Take the square root of the diagonal elements to obtain the standard errors.

Maximum Likelihood Estimation

Cov $(\hat{\boldsymbol{\theta}})$. Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimate.

Equation for (asymptotic) standard errors, square root of the diagonal of the inverse of the negative Hessian matrix:

$$\left[-\frac{\partial^2 \log L(\boldsymbol{\theta}; \text{data})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \Big|_{\hat{\boldsymbol{\theta}}} \right]^{-1},$$

that is, get the Hessian matrix of negative second order derivatives, take the inverse, extract the diagonal components and take square roots.

The Hessian of g measures the curvature of the negative log-likelihood surface at $\hat{\boldsymbol{\theta}}$. The sharper the curvature is, the smaller the "uncertainty" and the smaller $\pm$ figure for the SE.

The more curved the surface (or parabola if $\boldsymbol{\theta}$ has dimension 1), the larger the Hessian (second derivative) and the smaller the inverse Hessian. SEs come from the sqrt of the diagonal elements of the inverse Hessian.

Maximum Likelihood Estimation

To get standard errors, confidence intervals, we must get the second derivative of **the logistic negative log-likelihood:**

Let $\pi_i = \pi_i(\mathbf{x}_i; \boldsymbol{\beta}) = \exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}/[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}] = 1/[1+\exp\{-\mathbf{x}_i^T\boldsymbol{\beta}\}]$, $1-\pi_i = 1/[1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}]$. The gradient vector is:

$$-\frac{\partial \log L(\boldsymbol{\beta}; \text{data})}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{n} \mathbf{x}_i y_i + \sum_{i=1}^{n} \mathbf{x}_i \frac{\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i^T\boldsymbol{\beta}\}} = -\sum_{i=1}^{n} \mathbf{x}_i y_i + \sum_{i=1}^{n} \mathbf{x}_i \pi_i$$

The Hessian matrix of second order derivatives is:

$$-\frac{\partial^2 \log L(\boldsymbol{\beta}; \text{data})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \pi_i(1-\pi_i), \tag{1}$$

making use of $\frac{d}{dz}\frac{z}{1+z} = \frac{1}{(1+z)^2}$ and $dz/d\boldsymbol{\beta}^T = \mathbf{x}^T z$, $z = \exp\{\mathbf{x}^T\boldsymbol{\beta}\}$.

When (1) is evaluated at the maximum likelihood estimate, π_i is replaced by $\hat{\pi}_i = 1/[1+\exp\{-\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}]$. To check that this is valid, try to get this result in non-matrix form when $p = 1$ (one explanatory variable).

Diagnostics for logistic regression: **The Deviance**.

The deviance of a model is defined as equal to twice the negative log-likelihood:

Deviance(model) = -2 (log-Likelihood(model))

We often compare the **Null Deviance**, the Deviance for the null model (without X):

$$\text{Null Deviance} = -2 \cdot \sum_{i=1}^{n} (y_i log(\bar{y}) + (1 - y_i) log(1 - \bar{y}))$$

To the **Residual Deviance** (Deviance of model with covariates X):

$$\text{Residual Deviance} = -2 \cdot \sum_{i=1}^{n} (y_i log(\hat{\pi}_i) + (1 - y_i) log(1 - \hat{\pi}_i))$$

where:

$$\hat{\pi}_i = Pr(Y = 1 | \mathbf{X} = x_i) = \frac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}}}$$

Lower **residual deviance** suggests that the model is better with the variables included.

Diagnostics for logistic regression: **The AIC**

1. The AIC seeks to identify "good models" by considering both the likelihood and the number of parameters in the model:

$$AIC = -2 \cdot log(Likelihood) + 2 \cdot (\#\text{ of parameters})$$

2. As such it is similar to the adjusted R^2.

3. Lower **AIC** suggests that the model is better.

Misclassification and the ROC curve

For comparison of logistic regression models with different subsets of explanatory variables, misclassification rates are one criteria; these can be in-sample (same data used to fit the data and estimate misclassification) or out-of-sample (training set to fit models and holdout set to estimate misclassification).

One can define

$$\hat{\pi}_i = \frac{\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}}{1+\exp\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}\}} \in [0,1], \quad \mathbf{x}_i^T \text{ is row } i \text{ of } \mathbf{X}, \quad i = 1,\ldots,n, \tag{6.7}$$

One can then set a threshold $\tau \in (0,1)$ and define a predicted $\hat{y}_i$, where

$$\hat{y}_i = \begin{cases} 0 & \text{if } 0 \leq \hat{\pi}_i \leq \tau, \\ 1 & \text{if } \tau < \hat{\pi}_i \leq 1. \end{cases} \tag{6.8}$$

The choice of threshold depends on the distribution of the n $\hat{\pi}_i$'s. If the proportion of 1s in the y_i is closer to 0, then so is the median of the $\hat{\pi}_i$'s. The counts and proportions for the in-sample misclassification are summarized in the table below. There are two types of misclassification errors: (i) false positive: predicting $\hat{y} = 1$ when true class is $y = 0$; (ii) false negative: predicting $\hat{y} = 0$ when true class is $y = 1$.

Misclassification and the ROC curve

One can then set a threshold $\tau \in (0,1)$ and define a predicted $\hat{y}_i$, where

$$\hat{y}_i = \begin{cases} 0 & \text{if } 0 \leq \hat{\pi}_i \leq \tau, \\ 1 & \text{if } \tau < \hat{\pi}_i \leq 1. \end{cases} \tag{6.8}$$

The choice of threshold depends on the distribution of the n $\hat{\pi}_i$'s. If the proportion of 1s in the y_i is closer to 0, then so is the median of the $\hat{\pi}_i$'s. The counts and proportions for the in-sample misclassification are summarized in the table below. There are two types of misclassification errors: (i) false positive: predicting $\hat{y} = 1$ when true class is $y = 0$; (ii) false negative: predicting $\hat{y} = 0$ when true class is $y = 1$.

	$\hat{\pi} \leq \tau$	$\hat{\pi} > \tau$	count	misclass rate
$y = 0$	n_{00}	n_{01}	$n_{00} + n_{01}$	$n_{01}/[n_{00} + n_{01}]$
$y = 1$	n_{10}	n_{11}	$n_{10} + n_{11}$	$n_{10}/[n_{10} + n_{11}]$
all	$n_{00} + n_{10}$	$n_{01} + n_{11}$	n	$(n_{01} + n_{10})/n$

Misclassification and the ROC curve

One can then set a threshold $\tau \in (0,1)$ and define a predicted $\hat{y}_i$, where

$$\hat{y}_i = \begin{cases} 0 & \text{if } 0 \leq \hat{\pi}_i \leq \tau, \\ 1 & \text{if } \tau < \hat{\pi}_i \leq 1. \end{cases} \tag{6.8}$$

The choice of threshold depends on the distribution of the n $\hat{\pi}_i$'s. If the proportion of 1s in the y_i is closer to 0, then so is the median of the $\hat{\pi}_i$'s. The counts and proportions for the in-sample misclassification are summarized in the table below. There are two types of misclassification errors: (i) false positive: predicting $\hat{y} = 1$ when true class is $y = 0$; (ii) false negative: predicting $\hat{y} = 0$ when true class is $y = 1$.

	$\hat{\pi} \leq \tau$	$\hat{\pi} > \tau$	count	misclass rate
$y = 0$	n_{00}	n_{01}	$n_{00} + n_{01}$	$n_{01}/[n_{00} + n_{01}]$
$y = 1$	n_{10}	n_{11}	$n_{10} + n_{11}$	$n_{10}/[n_{10} + n_{11}]$
all	$n_{00} + n_{10}$	$n_{01} + n_{11}$	n	$(n_{01} + n_{10})/n$

True Negatives

False Negatives

True Positives

False Positives

Misclassification and the ROC curve

Note that: The misclassification rate among the true 0s is $n_{01}/[n_{00}+n_{01}]$ and this decreases as τ increases. The misclassification rate among the true 1s is $n_{10}/[n_{10}+n_{11}]$ and this increases as τ increases.

Sensitivity: True Positive rate (= n11/(n11+n10))

Specificity: True Negative rate (= n00/(n00 + n01))

	$\hat{\pi} \leq \tau$	$\hat{\pi} > \tau$	count	misclass rate
$y = 0$	n_{00}	n_{01}	$n_{00}+n_{01}$	$n_{01}/[n_{00}+n_{01}]$
$y = 1$	n_{10}	n_{11}	$n_{10}+n_{11}$	$n_{10}/[n_{10}+n_{11}]$
all	$n_{00}+n_{10}$	$n_{01}+n_{11}$	n	$(n_{01}+n_{10})/n$

True Negatives

False Negatives

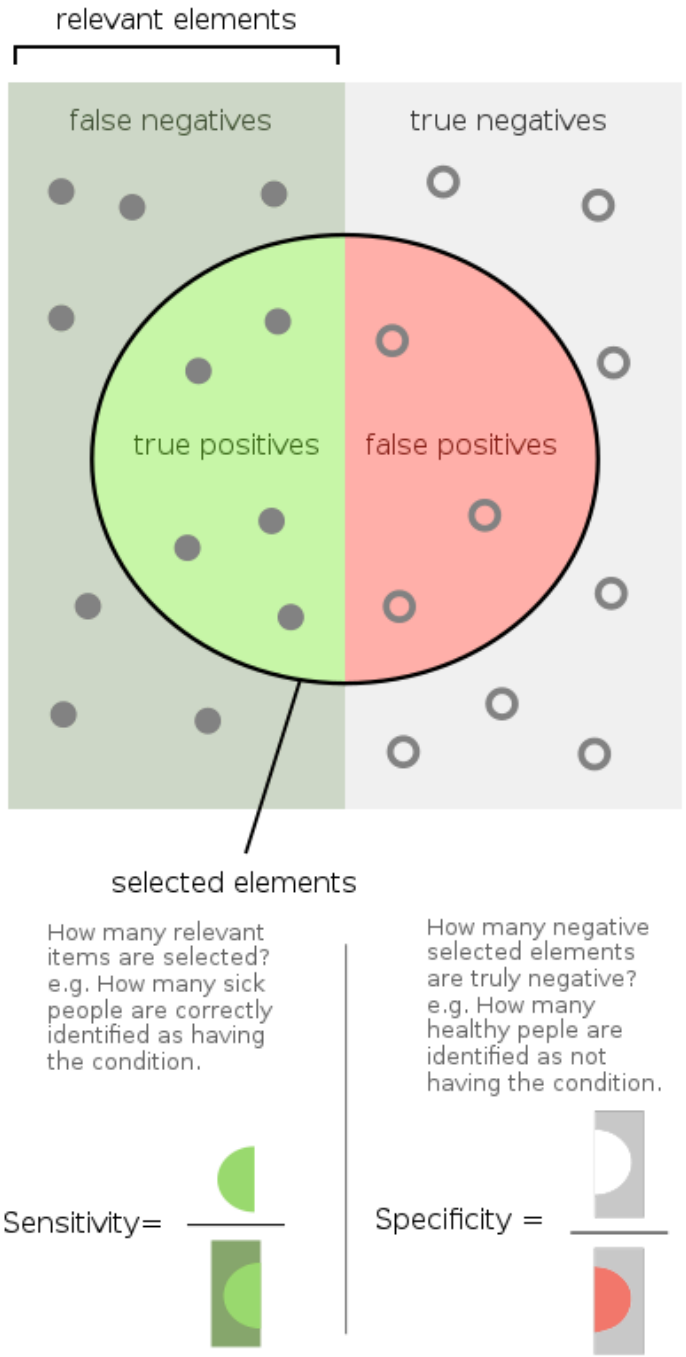
True Positives

False Positives

Misclassification and the ROC curve

Sensitivity: True Positive rate (= n11/(n11+n10))

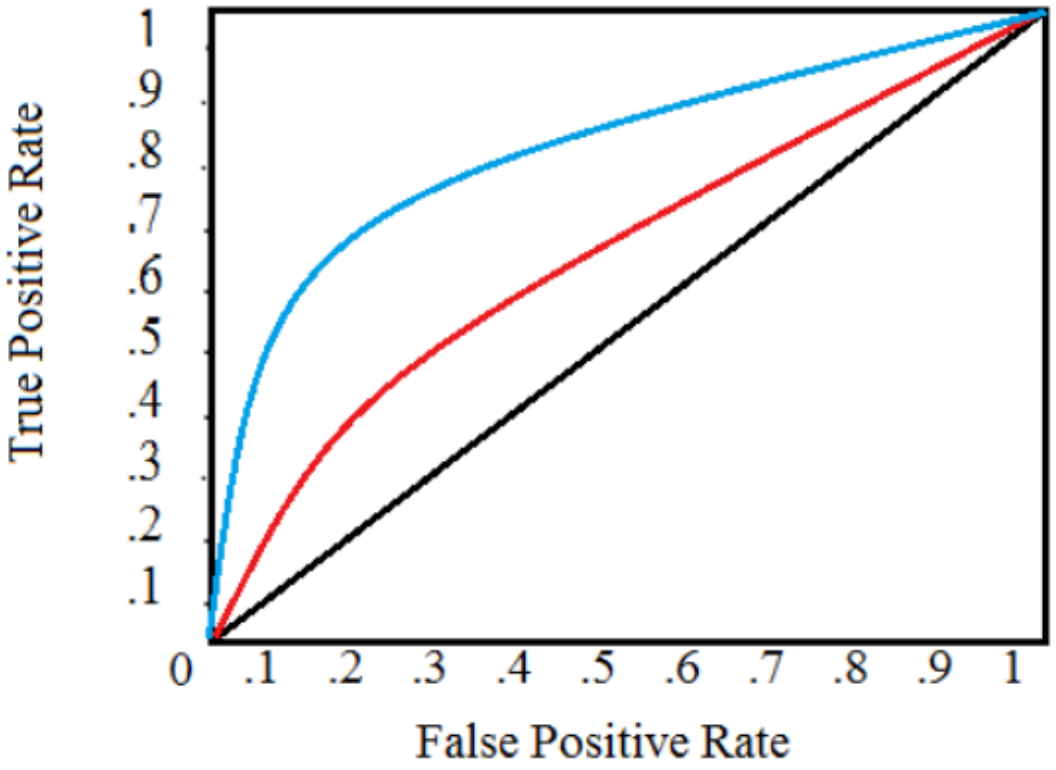
Specificity: True Negative rate (= n00/(n00 + n01))



Source: https://commons.wikimedia.org/w/index.php?curid=65826093

The Receiver Operating Characteristic curve (ROC curve)

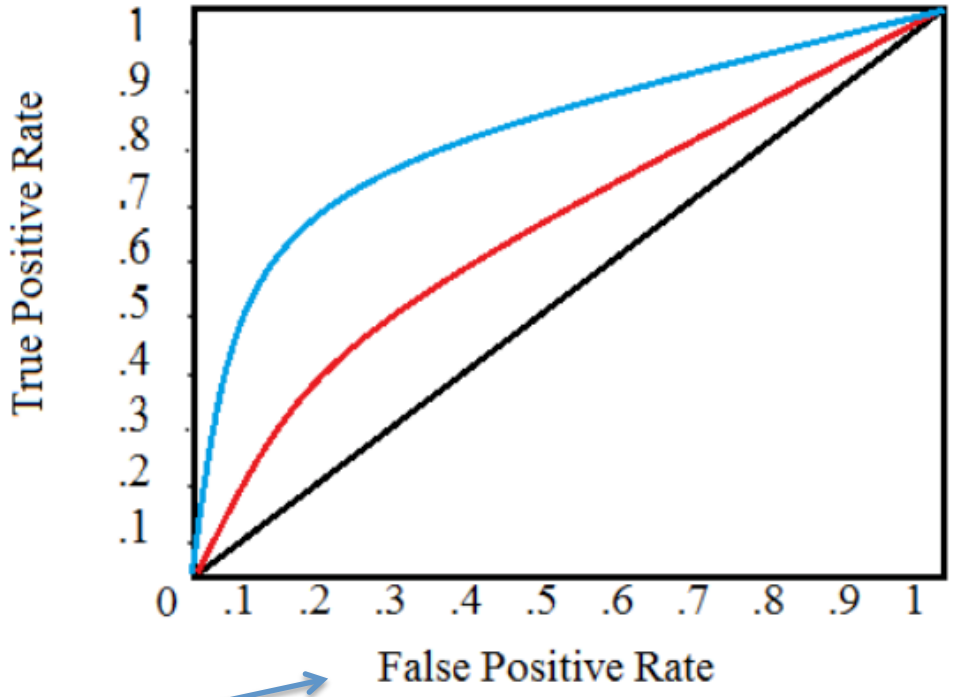
Is a plot that show how **Sensitivity** and **Specificity** change with different values for the threshold:



A ROC curve showing two tests. The red test is closer to the diagonal and is therefore less accurate than the green test.

The Receiver Operating Characteristic curve (ROC curve)

The ROC curve is a plot that show how **Sensitivity** and **Specificity** change with different values for the threshold:



Sensitivity

1- Specificity

A ROC curve showing two tests. The red test is closer to the diagonal and is therefore less accurate than the green test.

http://www.statisticshowto.com/receiver-operating-characteristic-roc-curve/

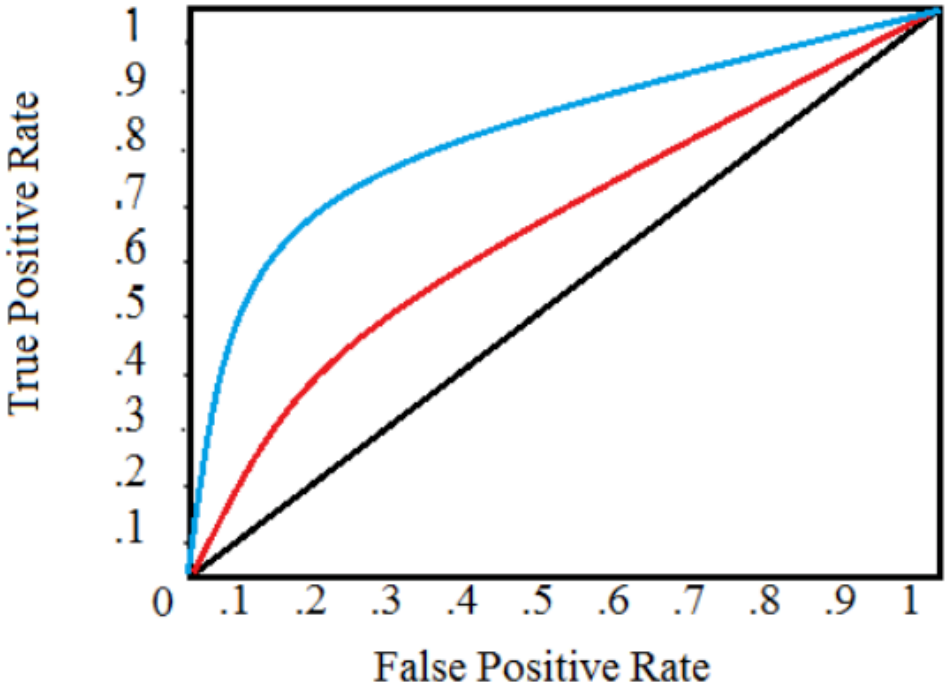
The Receiver Operating Characteristic curve (ROC curve)

The ROC curve is a plot that show how **Sensitivity** and **Specificity** change with different values for the threshold:

We can measure how good a model is by looking at the area under the ROC curve, also known as the AUC.

AUC will be between 0.5 and 1.

Higher AUC values indicate better prediction ability.



A ROC curve showing two tests. The red test is closer to the diagonal and is therefore less accurate than the green test.

http://www.statisticshowto.com/receiver-operating-characteristic-roc-curve/

Questions?

Excellent explanation of MLE for Logistic regression:
http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf

Principal component analysis.

- Principal component analysis (PCA) is a dimensionality reduction technique
  - many statistical models suffer from high correlation between covariates.
  - PCA can be used to produce linear combinations of the covariates that are uncorrelated between each other.
  - Often there are too many potential variables.
  - Sometimes there are more variables than observations (p>n)!
  - PCA can be used to reduce the number of variables in the model while maintaining as much information as possible.

Excellent explanation:

http://www.milanor.net/blog/performing-principal-components-regression-pcr-in-r/

Steps to obtain principal components:

1. Calculate the estimated covariance matrix of X (N observations, K variables), is the matrix Σ with entries:

$$\Sigma_{jk} = \frac{1}{N-1} \sum_{i=1}^{N} \left(X_{ij} - \bar{X}_j\right) \left(X_{ik} - \bar{X}_k\right)$$ for k in 1,…K, and j in 1,...,J

https://en.wikipedia.org/wiki/Covariance#Calculating_the_sample_covariance

2. Obtain eigenvalues and their corresponding eigenvectors for this covariance matrix.

https://www.youtube.com/watch?v=IdsV0RaC9jM&ab_channel=patrickJMT

3. Reduce X to a lower-dimensional projection: the eigenvectors corresponding to the largest eigenvalues.

4. Center the eigenvectors. This is the "best" projection of X onto the lower-dimension.

5. Determine what percentage of the original variance is maintained in the lower-dimensional projection.

Why does this work?

Suppose $\mathbf{X} = (X_1, \ldots, X_p)^T$ is a random vector with covariance matrix $\mathbf{\Sigma} = (\sigma_{jk})_{1 \le j,k \le p}$. Data are $(x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, considered as n independent realizations of $(X_1, \ldots, X_p)$. The sample covariance matrix is denoted as $\mathbf{S}$, and this is an estimate of $\mathbf{\Sigma}$.

From Appendix A, $\mathbf{w}^T\mathbf{X} = \sum_{j=1}^{p} w_j X_j$ is a linear combination with variance $\mathrm{Var}\,(\mathbf{w}^T\mathbf{X}) = \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} = \sum_{j=1}^{p}\sum_{k=1}^{p} w_j w_k \sigma_{jk}$.

For the *first principal component*, we want to find a coefficient vector $\mathbf{w}$ with unit length such that

$$\mathrm{Var}\,(\mathbf{w}^T\mathbf{X}) = \mathrm{Var}\,(\mathbf{w}^T(\mathbf{X} - \mathbf{a})) = \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} \tag{7.1}$$

is maximized. Why is unit length $\mathbf{w}^T\mathbf{w} = 1$ condition imposed?

Why does this work?

To solve this problem using calculus, we introduce the Lagrangian function:

$$H(\mathbf{w}) = \mathbf{w}^T \boldsymbol{\Sigma} \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{w} - 1), \tag{7.2}$$

$$\frac{\partial H}{\partial \mathbf{w}} = 2\boldsymbol{\Sigma}\mathbf{w} - 2\lambda\mathbf{w}. \tag{7.3}$$

Set to zero and let the solution involve $(\mathbf{w}^*, \lambda^*)$:

$$2\boldsymbol{\Sigma}\mathbf{w}^* - 2\lambda^*\mathbf{w}^* = 0, \text{ or } \boldsymbol{\Sigma}\mathbf{w}^* = \lambda^*\mathbf{w}^*. \tag{7.4}$$

Equation (7.4) is the equation for an eigensystem (review your linear algebra textbook if needed). So $\mathbf{w}^*$ is a unit-length eigenvector of $\boldsymbol{\Sigma}$ with eigenvalue λ^*. Because $\boldsymbol{\Sigma}$ is symmetric, the eigenvalues are real. Because $\boldsymbol{\Sigma}$ is non-negative definite, the eigenvalues are non-negative.

However….

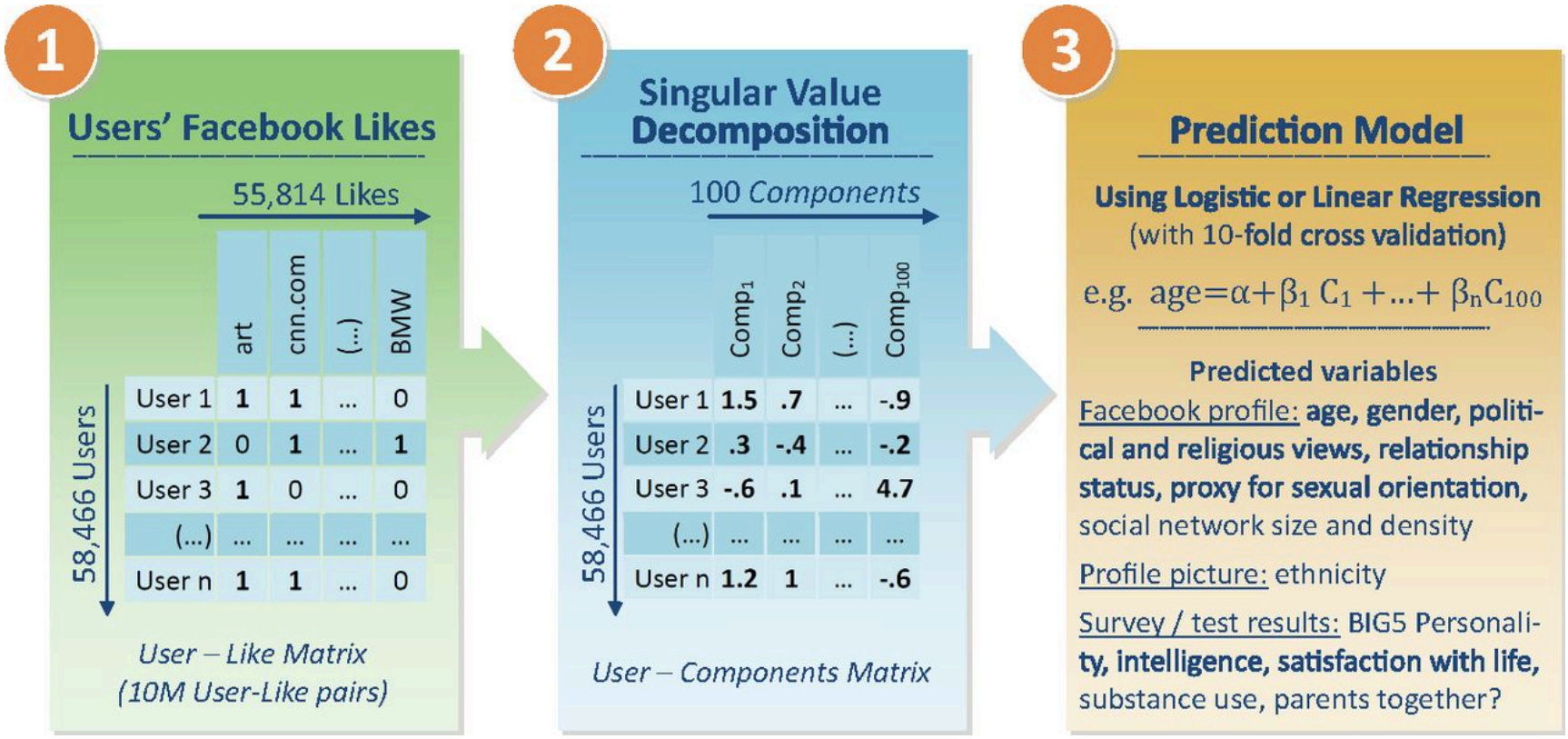
Instead of using the covariance matrix of X to do PCA, we should use the correlation matrix of X. In this way all the variables in the X matrix are on a similar scale.

To go from a Covariance matrix, Σ, to a correlation matrix ρ:

$$D = \sqrt{diag(\Sigma)}$$

$$\rho = D^{-1}\Sigma D^{-1}$$

Then obtain eigenvalues and their corresponding eigenvectors for ρ as before.


1
Users' Facebook Likes
55,814 Likes
58,466 Users
art
cnn.com
(...)
BMW
User 1 1 1 ... 0
User 2 0 1 ... 1
User 3 1 0 ... 0
(...) ... ...
User n 1 1 ... 0
User – Like Matrix
(10M User-Like pairs)
2
Singular Value Decomposition
100 Components
58,466 Users
Comp1
Comp2
(...)
Comp100
User 1 1.5 .7 ... -.9
User 2 .3 -.4 ... -.2
User 3 -.6 .1 ... 4.7
(...) ... ...
User n 1.2 1 ... -.6
User – Components Matrix
3
Prediction Model
Using Logistic or Linear Regression
(with 10-fold cross validation)
e.g. age=α+β1 C1 +...+ βnC100
Predicted variables
Facebook profile: age, gender, political and religious views, relationship status, proxy for sexual orientation, social network size and density
Profile picture: ethnicity
Survey / test results: BIG5 Personality, intelligence, satisfaction with life, substance use, parents together?