Announcements

- Google form feedback https://tinyurl.com/yb2tprkl

The previous and future weeks

Machine Learning – CSE546

Kevin Jamieson
University of Washington

November 9, 2017

So far…

Supervised learning: $x_i \in R^d$ $y_i \in \mathbb{R}$ for $i = 1, \ldots, n$. Learn $f : x \to y$

Loss functions: $L(f) = \sum_{i=1} \ell(f(x_i), y_i)$

$\ell(y, f(x)) = (f(x)-y)^2$

$\max(0, 1-yf(x))$

$= y\log(f(x)) + (1-y)\log(1-f(x)) \Longleftrightarrow \log(1+\exp(-yf(x)))$

$e^{yf(x)}$

Methods:

Linear
- Lasso
- Ridge

SVM

Trees, Boosting Trees

Nearest Neighbor

Kernel Machines

MLE, MAP

Neural Networks

Bagging

Bias-Variance

Overfitting

Cross Validation

Model Assessment (Hoeffding)

Method comparison

TABLE 10.1. *Some characteristics of different learning methods. Key:* ▲ *= good,* ◆ *=fair, and* ▼ *=poor.*

Characteristic	Neural Nets	SVM	Trees	Boosting Trees	k-NN, Kernels
Natural handling of data of "mixed" type	▼	▼	▲	▲	▼
Handling of missing values	▼	▼	▲	▲	▲
Robustness to outliers in input space	▼	▼	▲	▼	▲
Insensitive to monotone transformations of inputs	▼	▼	▲	▼	▼
Computational scalability (large N)	▼	▼	▲	▲	▼
Ability to deal with irrelevant inputs	▼	▼	▲	▲	▼
Ability to extract linear combinations of features	▲	▲	▼	▼	◆
Interpretability	▼	▼	◆	▲	▼
Predictive power	▲	▲	▼	◆	▲

©Kevin Jamieson 2017

To come

- Unsupervised learning
  - SVD
  - Clustering
  - Density estimation
- Machine learning street fighting tools
  - Tips, tricks, data pre-processing, output post-processing
  - Domain specific data (images, sequences)
- Reinforcement learning
- Learning theory

Principle Component Analysis

Machine Learning – CSE546
Kevin Jamieson
University of Washington

November 9, 2017

Linear projections

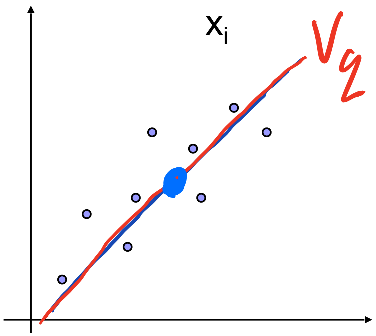
Given $x_i \in \mathbb{R}^d$ and some $q < d$ consider

$$\min_{\mu, \{\lambda_i\}, \mathbf{V}_q} \sum_{i=1}^{N} \|x_i - \mu - \mathbf{V}_q \lambda_i\|^2.$$

$V_q \in \mathbb{R}^{d \times q}$

where $\lambda_i \in \mathbb{R}^q$ and $\mathbf{V}_q = [v_1, v_2, \ldots, v_q]$ is orthonormal:

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$



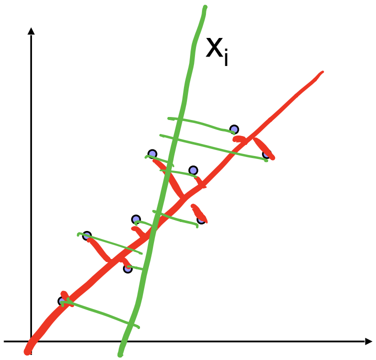
Linear projections

Given $x_i \in \mathbb{R}^d$ and some $q < d$ consider

$$\min_{\mu, \{\lambda_i\}, \mathbf{V}_q} \sum_{i=1}^{N} \|x_i - \underline{\mu} - \underbrace{\mathbf{V}_q \lambda_i}\|^2.$$

where $\lambda_i \in \mathbb{R}^q$ and $\mathbf{V}_q = [v_1, v_2, \ldots, v_q]$ is orthonormal:

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$

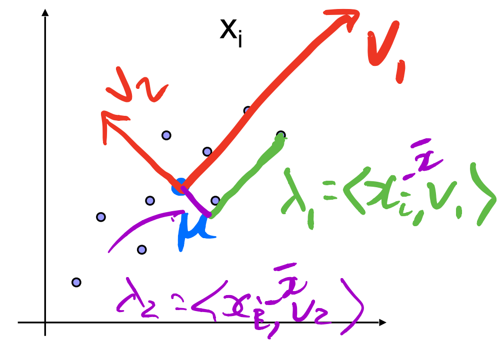


Natural choices for μ, λ_i ?

Linear projections

Given $x_i \in \mathbb{R}^d$ and some $q < d$ consider

$$\min_{\mu,\{\lambda_i\},\mathbf{V}_q} \sum_{i=1}^{N} \|x_i - \mu - \mathbf{V}_q \lambda_i\|^2.$$



where $\lambda_i \in \mathbb{R}^q$ and $\mathbf{V}_q = [v_1, v_2, \ldots, v_q]$ is orthonormal:

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$

$$x_i = \bar{x} + \lambda_i v_i$$

Natural choices for μ, λ_i ?

$$\begin{aligned} \hat{\mu} &= \bar{x}, \quad = \frac{1}{n} \sum_{i=1}^{n} x_i \\ \hat{\lambda}_i &= \mathbf{V}_q^T (x_i - \bar{x}). \end{aligned}$$

Which gives us:

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||^2.$$

$\mathbf{V}_q \mathbf{V}_q^T$ is a *projection matrix* that minimizes error in basis of size q

Linear projections

$Tr(v^T A v) = Tr(A v v^T)$

$Tr((A+B)C) = Tr(AC) + Tr(BC)$

$= I - 2V_q V_q^T + V_q V_q^T V_q V_q^T$

$= I - V_q V_q^T$

$$\Sigma := \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T$$

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$

$$\sum_{i=1}^N ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||_2^2$$

$$= \sum || (I - V_q V_q^T)(x_i - \bar{x})||_2^2$$

$$= \sum_i (x_i - \bar{x})^T (I - V_q V_q^T)^T (I - V_q V_q^T)(x_i - \bar{x})$$

$$= \sum_i (x_i - \bar{x})^T (I - V_q V_q^T)(x_i - \bar{x})$$

$$= \sum_i Tr((I - V_q V_q^T)(x_i - \bar{x})(x_i - \bar{x})^T)$$

$$= \sum Tr((x_i - \bar{x})(x_i - \bar{x})^T) - Tr(V_q V_q^T (x_i - \bar{x})(x_i - \bar{x})^T)$$

$$= Tr(\Sigma) - \sum_{i=1}^n Tr(V_q^T (x_i - \bar{x})(x_i - \bar{x})^T V_q)$$

$$= Tr(\Sigma) - Tr(V_q^T \Sigma V_q)$$

Linear projections

$$\sum_{i=1}^N ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||_2^2$$

$$\Sigma := \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T$$

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$

λ, v are eigenvalue/vector pair if

$\Sigma v = \lambda v$

$$\tilde{\Lambda} = \begin{bmatrix} \lambda_1 & & 0 \\ 0 & \lambda_2 & \\ & & \ddots \ \lambda_d \end{bmatrix}$$

$$\min_{\mathbf{V}_q} \sum_{i=1}^N ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||_2^2 = \min_{\mathbf{V}_q} Tr(\Sigma) - \underline{Tr(\mathbf{V}_q^T \Sigma \mathbf{V}_q)}$$

$\Sigma = \tilde{V} \tilde{\Lambda} \tilde{V}^T$

$\max_{V} Tr(V_q^T \tilde{V} \tilde{\Lambda} \tilde{V}^T V_q)$

$\Longrightarrow V_q = \tilde{V}_q$

Linear projections

$$\sum_{i=1}^N ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||_2^2$$

$$\Sigma := \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T$$

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$

$$\min_{\mathbf{V}_q} \sum_{i=1}^N ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||_2^2 = \min_{\mathbf{V}_q} Tr(\Sigma) - Tr(\mathbf{V}_q^T \Sigma \mathbf{V}_q)$$

Eigenvalue decomposition of Σ

$\mathbf{V}_q$ are the first q eigenvectors of Σ

with the largest q eigenvalues

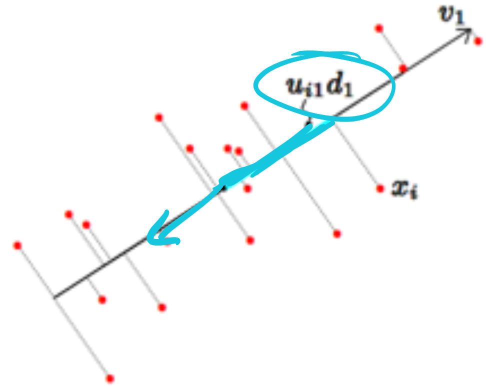
Linear projections

Given $x_i \in \mathbb{R}^d$ and some $q < d$ consider

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||^2.$$

where $\mathbf{V}_q = [v_1, v_2, \dots, v_q]$ is orthonormal:

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$



$\mathbf{V}_q$ are the first q eigenvectors of Σ

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

$\mathbf{V}_q$ are the first q *principle components*

Principle component Analysis (PCA) projects $(\mathbf{X} - \mathbf{1}\bar{x})$ down onto $\mathbf{V}_q$

$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \mathbf{U}_q \mathrm{diag}(d_1, \dots, d_q) \qquad \mathbf{U}_q^T \mathbf{U}_q = I_q$$

Linear projections

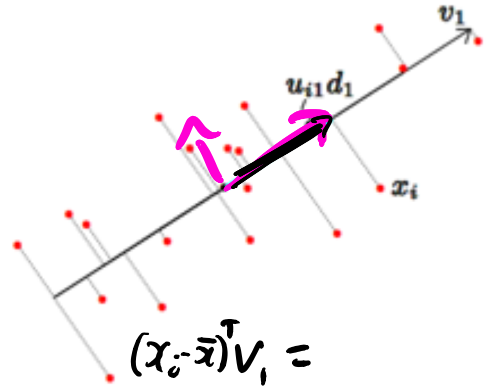
$\bar{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$

Given $x_i \in \mathbb{R}^d$ and some $q < d$ consider

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - \underline{\mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})}||^2.$$

where $\mathbf{V}_q = [v_1, v_2, \ldots, v_q]$ is orthonormal:

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$



$(x_i - \bar{x})^T v_1 =$

$\mathbf{V}_q$ are the first q eigenvectors of Σ

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

$\mathbf{V}_q$ are the first q *principle components*

Principle component Analysis (PCA) projects $(\mathbf{X} - \mathbf{1}\bar{x})$ down onto $\mathbf{V}_q$ (if $d < n$)

$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \mathbf{U}_q \text{diag}(d_1, \ldots, d_q) \qquad \mathbf{U}_q^T \mathbf{U}_q = I_q$$

Singular Value Decomposition defined as

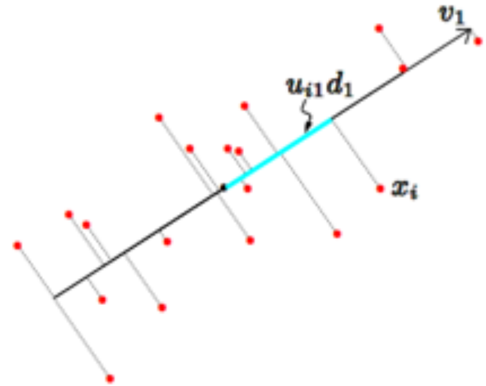
$$\underline{\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{USV}^T}$$

$U \in \mathbb{R}^{n \times d}$
$S \in \mathbb{R}^{d \times d}$
$V = \mathbb{R}^{d \times d}$

Linear projections

Given $x_i \in \mathbb{R}^d$ and some $q < d$ consider

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||^2.$$



where $\mathbf{V}_q = [v_1, v_2, \dots, v_q]$ is orthonormal:

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$

$\mathbf{V}_q$ are the first q eigenvectors of Σ

$\mathbf{V}_q$ are the first q *principle components*

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T = (X - 1\bar{x}^T)^T (X - 1\bar{x}^T)$$

Principle component Analysis (PCA) projects $(\mathbf{X} - \mathbf{1}\bar{x})$ down onto $\mathbf{V}_q$

$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \underline{\mathbf{U}_q \text{drag}(d_1, \dots, d_q)} \qquad \mathbf{U}_q^T \mathbf{U}_q = I_q$$

$$\Sigma = \sum_{i=1}^{d} v_i v_i^T d_i^2$$

Singular Value Decomposition defined as

$$\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

How do the *eigenvalues* of Σ relate to the *singular values* of $\mathbf{X} - \mathbf{1}\bar{x}$?

A is singular if $\exists x \neq 0 : Ax = 0$

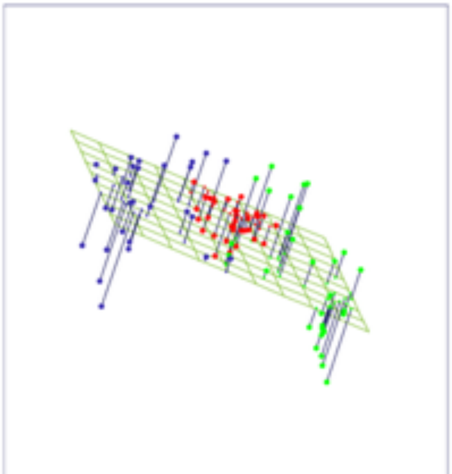
Dimensionality reduction

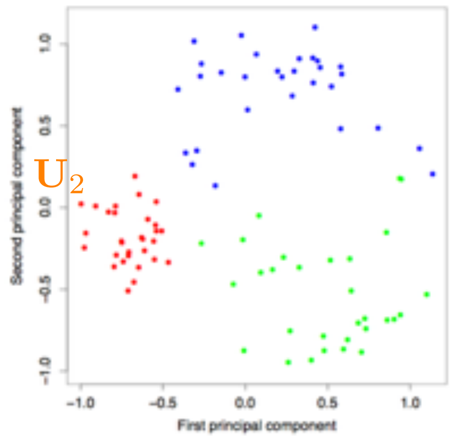
$\mathbf{V}_q$ are the first q eigenvectors of Σ and *SVD* $\underline{\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{USV}^T}$

singular values of $X - \mathbf{1}\bar{x}^T$

$(AB)^T = B^T A^T$

$X \in \mathbb{R}^{n \times d}$





$\mathbf{U}_2$

$\mathbf{U}_1$

$\mathbf{X} - \mathbf{1}\bar{x}^T$

eigenvalues of Σ are singular values squared.

eigenvalues of Σ?

$$\Sigma = (X - \mathbf{1}\bar{x}^T)^T (X - \mathbf{1}\bar{x}^T) = V \underbrace{S^T}_{=S} \underbrace{U^T U}_{I} S V^T = V S^2 V^T$$

Dimensionality reduction

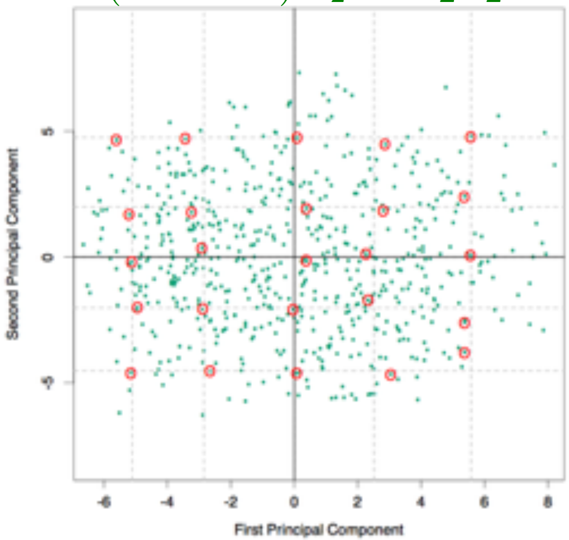
$\mathbf{V}_q$ are the first q eigenvectors of Σ and *SVD* $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$

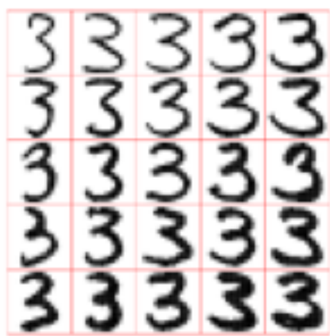
Handwritten 3's, 16x16 pixel image so that $x_i \in \mathbb{R}^{256}$

$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$

$=$  $+ \lambda_1 \cdot$  $+ \lambda_2 \cdot$ .

$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_2 = \mathbf{U_2 S_2} \in \mathbb{R}^{n \times 2}$





diag($\mathbf{S}$)

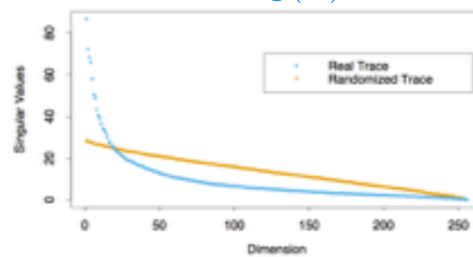


FIGURE 14.24. *The 256 singular values for the digitized threes, compared to those for a randomized version of the data (each column of* $\mathbf{X}$ *was scrambled).*

Kernel PCA

$\mathbf{V}_q$ are the first q eigenvectors of Σ and *SVD* $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{USV}^T$

$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \mathbf{U_q S_q} \in \mathbb{R}^{n \times q}$$

$\bar{x}^T = \mathbf{1}^T X / n$

$$\mathbf{JX} = \mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{USV}^T \qquad \mathbf{J} = I - \mathbf{1}\mathbf{1}^T/n$$

$$(\mathbf{JX})(\mathbf{JX})^T = J\underbrace{XX^T}_{K}J = (X - \mathbf{1}\bar{x}^T)(X - \mathbf{1}\bar{x}^T)^T$$

$$= USV^T VSU^T$$

$$= US^2U^T$$

Kernel PCA

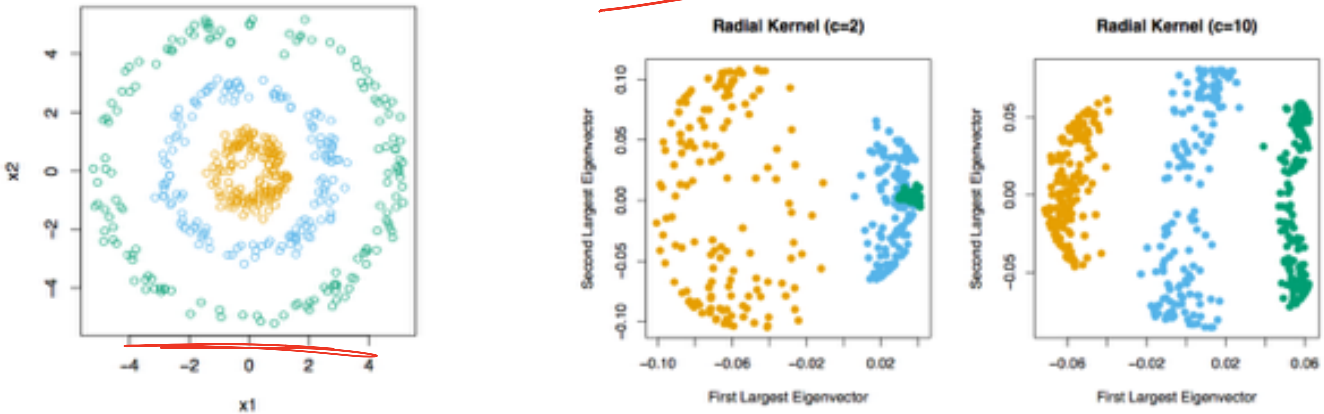
$\mathbf{V}_q$ are the first q eigenvectors of Σ and *SVD* $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$

$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \mathbf{U_q S_q} \in \mathbb{R}^{n \times q}$$

$$\mathbf{J}\mathbf{X} = \mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T \qquad \mathbf{J} = I - \mathbf{1}\mathbf{1}^T/n$$

$$K_{ij} = e^{-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}}$$

$$(\mathbf{J}\mathbf{X})(\mathbf{J}\mathbf{X})^T = \mathbf{U}\mathbf{S}^2\mathbf{U}^T = \underline{JKJ}$$



PCA Algorithm

PCA

input
A matrix of m examples $X \in \mathbb{R}^{m,d}$
number of components n
if $(m > d)$
$A = X^\top X$
Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be the eigenvectors of A with largest eigenvalues
else
$B = X X^\top$
Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be the eigenvectors of B with largest eigenvalues
for $i = 1, \ldots, n$ set $\mathbf{u}_i = \frac{1}{\|X^\top \mathbf{v}_i\|} X^\top \mathbf{v}_i$
output: $\mathbf{u}_1, \ldots, \mathbf{u}_n$

©Kevin Jamieson 2017

Ridge Regression revisited

$(AB)^{-1} = B^{-1}A^{-1}$ $\quad V^TV = I$ $\quad V^{-1}V = I$

$V^{-1} = V^T$

$$\widehat{w}_{ridge} = \arg\min_w ||\mathbf{X}w - \mathbf{y}||_2^2 + \lambda||w||_2^2$$

$$\widehat{w}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

Assume X is centered

Singular vector decomposition (SVD): $\mathbf{X} \cancel{= \mathbf{U}\Sigma\mathbf{V}^T} = \mathbf{USV}^T$

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

$$= USV^T(VS^2V^T + \lambda I)^{-1}VSU^Ty$$

$U = [u_1, \ldots u_d]$

$$= USV^T(V(S^2+\lambda I)V^T)^{-1}VSU^T$$

$$= USV^T\left(V^{T^{-1}}(S^2+\lambda I)^{-1}V^{-1}\right)VSU^T$$

$$= USV^TV(S^2+\lambda I)^{-1}V^TVSU^T$$

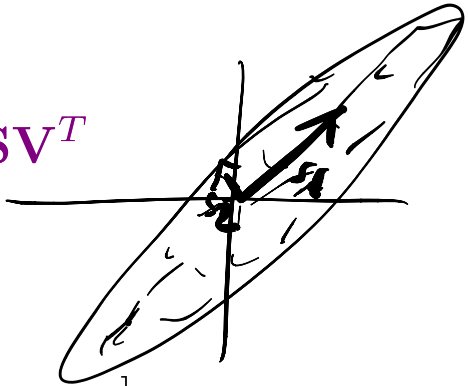
$$= US(S^2+\lambda I)^{-1}SU^T = \sum_{i=1}^{d} u_iu_i^T\frac{s_i^2}{s_i^2+\lambda}$$

Ridge Regression revisited

$$\widehat{w}_{ridge} = \arg\min_w ||\mathbf{X}w - \mathbf{y}||_2^2 + \lambda ||w||_2^2$$

$$\widehat{w}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

Singular vector decomposition (SVD): $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{USV}^T$



$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\mathbf{y}} = \sum_{i=1}^{d} u_i u_i^T \frac{s_i^2}{s_i^2 + \lambda} y_i$$

$$\mathbf{U} = [u_1, \dots, u_d]$$

$$\mathbf{S} = \mathrm{diag}(s_1, \dots, s_d)$$