Machine learning

• Unsupervised learning

• Supervised learning
Machine learning

• Unsupervised learning
  – dimension reduction, clustering
• Supervised learning
Machine learning

- Unsupervised learning
  - dimension reduction, clustering
- Supervised learning
  - classification, regression
Machine learning

• Unsupervised learning
  – dimension reduction, clustering
• Supervised learning
  – classification, regression
Principal Components Analysis (PCA)

- Dimension reduction
- Useful for exploratory data analysis of high-dimensional data sets.
Example: Consider a data set of heights and weights of people.
Example: Consider a data set of heights and weights of people.
Example: Consider a data set of heights and weights of people.
PCA on this data set reframes data in terms of overall size and heavyness.

heavyness = weight - height

overall size = weight + height

less heavy
smaller

heavier
bigger
The math behind PCA

Variance of one variable:

$$\text{Var}(X) = \frac{1}{n} \sum_j (\bar{x} - x_j)^2 = \sigma_x^2$$

Covariance of two variables:

$$\text{Cov}(X,Y) = \frac{1}{n} \sum_j (\bar{x} - x_j)(\bar{y} - y_j) = \sigma_{XY}^2$$
The math behind PCA

Covariance matrix of $n$ variables $X_1 \ldots X_n$:

$$
C = \begin{pmatrix}
\sigma^2_{11} & \sigma^2_{12} & \cdots & \sigma^2_{1n} \\
\sigma^2_{21} & \sigma^2_{22} & \cdots & \sigma^2_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
\sigma^2_{n1} & \sigma^2_{n2} & \cdots & \sigma^2_{nn}
\end{pmatrix}
$$
PCA diagonalizes the covariance matrix $C$:

$$ C = U D U^T $$

\[
U = \begin{pmatrix}
    \lambda_1^2 & 0 & \cdots & 0 \\
    0 & \lambda_2^2 & \cdots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \cdots & \lambda_n^2 \\
\end{pmatrix} U^T
\]
The math behind PCA

PCA diagonalizes the covariance matrix $C$:

$$C = \textbf{U} \textbf{D} \textbf{U}^T$$

where

$$\textbf{D} = \begin{pmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n^2 \end{pmatrix}$$

rotation matrix
The math behind PCA

PCA diagonalizes the covariance matrix $C$: 

$$C = U D U^T$$ 

$$= U \begin{pmatrix} 
\lambda_1^2 & 0 & \ldots & 0 \\
0 & \lambda_2^2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \lambda_n^2 
\end{pmatrix} U^T$$
The math behind PCA

PCA diagonalizes the covariance matrix \( C \):

\[
C = U \Lambda U^T
\]

where \( \Lambda \) is a diagonal matrix with eigenvalues \( \lambda_1^2, \lambda_2^2, \ldots, \lambda_n^2 \) on the diagonal. These eigenvalues represent the variance explained by each component.

Eigenvalues (\( \lambda_1^2, \lambda_2^2, \ldots, \lambda_n^2 \)) are the primary focus of PCA as they indicate the amount of variance in the data that each component captures.
The math behind PCA

PCA diagonalizes the covariance matrix $\mathbf{C}$:

$$\mathbf{C} = \mathbf{U} \mathbf{D} \mathbf{U}^T$$

$$= \mathbf{U} \begin{pmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n^2 \end{pmatrix} \mathbf{U}^T$$

covariance between components is zero (they are uncorrelated)
In our earlier example, overall size and heaviness are uncorrelated.
Doing a PCA in R

```r
iris %>%
  select(-Species) %>% # remove Species column
  scale() %>% # scale to zero mean
  # and unit variance
  prcomp() -> pca # do PCA
  # store result
  pca # in variable “pca”
```
Doing a PCA in R

> pca
Standard deviations:
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation:

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal.Length</td>
<td>0.521066</td>
<td>-0.37742</td>
<td>0.719566</td>
<td>0.261286</td>
</tr>
<tr>
<td>Sepal.Width</td>
<td>-0.26935</td>
<td>-0.92330</td>
<td>-0.24438</td>
<td>-0.12351</td>
</tr>
<tr>
<td>Petal.Length</td>
<td>0.580413</td>
<td>-0.02449</td>
<td>-0.14213</td>
<td>-0.80145</td>
</tr>
<tr>
<td>Petal.Width</td>
<td>0.56486</td>
<td>-0.06694</td>
<td>-0.63427</td>
<td>0.523597</td>
</tr>
</tbody>
</table>
Doing a PCA in R

```r
> pca

Standard deviations:
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation:

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal.Length</td>
<td>0.5210659</td>
<td>-0.37741762</td>
<td>0.7195664</td>
<td>0.2612863</td>
</tr>
<tr>
<td>Sepal.Width</td>
<td>-0.2693474</td>
<td>-0.92329566</td>
<td>-0.2443818</td>
<td>-0.1235096</td>
</tr>
<tr>
<td>Petal.Length</td>
<td>0.5804131</td>
<td>-0.02449161</td>
<td>-0.1421264</td>
<td>-0.8014492</td>
</tr>
<tr>
<td>Petal.Width</td>
<td>0.5648565</td>
<td>-0.06694199</td>
<td>-0.6342727</td>
<td>0.5235971</td>
</tr>
</tbody>
</table>
```
Squares of the std. devs represent the % variance explained by each PC.
Doing a PCA in R

```r
> pca
Standard deviations:
[1] 1.7083611 0.9560494 0.3830886 0.1439265

Rotation:

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal.Length</td>
<td>0.5210659</td>
<td>-0.37741762</td>
<td>0.7195664</td>
<td>0.2612863</td>
</tr>
<tr>
<td>Sepal.Width</td>
<td>-0.2693474</td>
<td>-0.92329566</td>
<td>-0.2443818</td>
<td>-0.1235096</td>
</tr>
<tr>
<td>Petal.Length</td>
<td>0.5804131</td>
<td>-0.02449161</td>
<td>-0.1421264</td>
<td>-0.8014492</td>
</tr>
<tr>
<td>Petal.Width</td>
<td>0.5648565</td>
<td>-0.06694199</td>
<td>-0.6342727</td>
<td>0.5235971</td>
</tr>
</tbody>
</table>
```
The rotation matrix tells us which variables contribute to which PCs.
We can also recover each original observation expressed in PC coordinates

```r
> pca$x
```
We can also recover each original observation expressed in PC coordinates.

```r
> pca$x

<table>
<thead>
<tr>
<th></th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2.2571</td>
<td>-0.4784</td>
<td>0.1273</td>
<td>0.0241</td>
</tr>
<tr>
<td>2</td>
<td>-2.0740</td>
<td>0.6719</td>
<td>0.2338</td>
<td>0.1026</td>
</tr>
<tr>
<td>3</td>
<td>-2.3563</td>
<td>0.3408</td>
<td>-0.0441</td>
<td>0.0282</td>
</tr>
<tr>
<td>4</td>
<td>-2.2917</td>
<td>0.5954</td>
<td>-0.091</td>
<td>-0.0657</td>
</tr>
<tr>
<td>5</td>
<td>-2.3818</td>
<td>-0.6447</td>
<td>-0.1568</td>
<td>-0.0358</td>
</tr>
<tr>
<td>6</td>
<td>-2.0687</td>
<td>-1.4842</td>
<td>-0.0268</td>
<td>0.0066</td>
</tr>
<tr>
<td>7</td>
<td>-2.4359</td>
<td>-0.0475</td>
<td>-0.3343</td>
<td>-0.0366</td>
</tr>
<tr>
<td>8</td>
<td>-2.2254</td>
<td>-0.2224</td>
<td>0.0884</td>
<td>-0.0245</td>
</tr>
<tr>
<td>9</td>
<td>-2.3268</td>
<td>1.1116</td>
<td>-0.1446</td>
<td>-0.0268</td>
</tr>
<tr>
<td>10</td>
<td>-2.1770</td>
<td>0.4674</td>
<td>0.2529</td>
<td>-0.0397</td>
</tr>
<tr>
<td>11</td>
<td>-2.1590</td>
<td>-1.0402</td>
<td>0.2678</td>
<td>0.0167</td>
</tr>
<tr>
<td>12</td>
<td>-2.3184</td>
<td>-0.1326</td>
<td>-0.0934</td>
<td>-0.1330</td>
</tr>
<tr>
<td>13</td>
<td>-2.2110</td>
<td>0.7262</td>
<td>0.2301</td>
<td>0.0024</td>
</tr>
</tbody>
</table>
Plot of iris plants in PC coordinates reveals differences among species

Species
- setosa
- versicolor
- virginica
These differences are much harder to see in the original variables.