Advanced Machine Learning Subsidary Notes

Lecture 7: Principal Component Analysis (PCA)

Adam Prügel-Bennett

April 21, 2020

1 Keywords

- Covariance matrices, dimensionality reduction, PCA, Duality

2 Main Points

2.1 PCA

- This is revision as you should have all seen this in foundations of ML
- The covariance matrix is defined as

$$\mathsf{C} = \frac{1}{m+1} \sum_{i=1}^{m} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})\,(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}})^{\mathsf{T}}$$

- Defining the matrix X as

$$\mathsf{X} = \frac{1}{\sqrt{m-1}} \big(\boldsymbol{x}_1 - \boldsymbol{\mu}, \boldsymbol{x}_2 - \boldsymbol{\mu}, \cdots \boldsymbol{x}_m - \boldsymbol{\mu}\big)$$

  then $\mathsf{C} = \mathsf{X}\,\mathsf{X}^{\mathsf{T}}$

- The *principal components* are the eigenvectors of the covariance matrix with the largest eigenvalues
- We can reduce the dimensionality of the inputs by projecting into the subspace spanned by the principal components
- We can reconstruct a vector from its principal component projection

$$\hat{\boldsymbol{x}} = \sum_i z_i\, \boldsymbol{v}_i$$

  - $\boldsymbol{v}_i$ are the principal components (eigenvectors of the covariance matrix with largest eigenvalues)
  - $z_i = \boldsymbol{v}_i^{\mathsf{T}} \boldsymbol{x}$ are the values of the new features
  - the sum is over the principal components that we

- The expected squared error reconstruction loss $\mathbb{E}\left[(\hat{\boldsymbol{x}} - \boldsymbol{x})^2\right]$ is equal to the sum of the eigenvalues we ignore

2.2 Duality

- We can define the dual matrix, $\mathbf{D} = \mathbf{X}^\mathsf{T}\mathbf{X}$ with components $D_{kl} = (\boldsymbol{x}_k - \boldsymbol{\mu})^\mathsf{T}(\boldsymbol{x}_k - \boldsymbol{\mu})$
- If $\boldsymbol{u}_i$ is and eigenvector of $\mathbf{D}$ with eigenvalue λ_i then $\boldsymbol{v}_i = \mathbf{D}\,\boldsymbol{u}_i$ is an eigenvector of $\mathbf{C}$ with the same eigenvalue
- Matrix $\mathbf{C}$ and $\mathbf{D}$ have exactly the same non-zero eigenvalues
- If we have more features than training example it is more efficient to work with $\mathbf{D}$ than $\mathbf{C}$
- Note in this case the training examples will not span the feature space. $\mathbf{D}$ describes the fluctuations in the space spanned by the examples

3 Exercises

3.1 Duality

- Show that if $\boldsymbol{u}_i$ is and eigenvector of $\mathbf{D}$ with eigenvalue λ_i then $\boldsymbol{v}_i = \mathbf{D}\,\boldsymbol{u}_i$ is an eigenvector of $\mathbf{C}$ with the same eigenvalue
- Answer in the lecture notes

4 Experiments

4.1 Duality

- Using Matlab/Octave of python illustrate that the dual matrix and covariance matrix have the same eigenvalues

```
X = randn(5,3)    % construct a random matrix
C = X*X'          % compute a sort of covariance matrix (haven't bothered removing mean
D = X'*X          % compute dual
[V,LC] = eig(C)   % compute eigensystem of C
[U,LD] = eig(D)   % compute eigensystem of D should have the same non-zero eigenvalues
u1 = X*U(:,1)     % left multiply and eigenvector or D by X
u1/norm(u1)       % normalise above should be the same as V(:,3) (could be V(:,4) or V(:,5))
```