UNIVERSITY OF CALIFORNIA, BERKELEY

Lecture 20 - CS: 189

Oscar Ortega

July 16, 2021

1 UNSUPERVISED LEARNING

We have sample points, but no labels! No classes, no y-values, nothing to predict. So our goal becomes to discover some structure in the data.

1.1 EXAMPLES

- Cluster: Partition data into groups of similar/nearby points.
- Dimensionality Reduction: data often lies near a low-dimensional subspace (or manifold) in feature space; Matrices have low-rank approximations.
- Density Estimation: Fit a continuous distribution to discrete data. eg: MLE

2 PRINCIPAL COMPONENTS ANALYSIS: (PCA) (KARL PEARSON, 1901)

Goal: given sample points in $\mathbb{R}^d$, find k directions that capture most of the variation.
Why?

- Find a small basis for representing variations in complex things: (eg - faces, genes)
- Reducing number of dimensions makes some computations cheaper, (eg - regression)
- remove irrelevant dimensions to reduce over-fitting in learning algs.
- like subset selection, but the "features" aren't axis-aligned. They're linear combos of input features.

Let $X \in \mathbb{R}^{n,d}$ assume X is centered, mean X_i is 0.

Let w be a unit vector. The orthogonal projection of x onto $w = (x^T w)$.

Recall if w is not a unit vector $\hat{x} = \frac{x^T w}{\|w\|_2^2} w$.

Given orthonormal directions $v_1, ..., v_k$: $\tilde{x} = \sum_k (x^T v_i)$

Recall that MLE estimates co-variance matrix $\Sigma = \frac{1}{n} X^T X$.

2.1 PCA - Alg:

- Center X
- Optional: Normalize X: Units of measurements different?
- Yes : normalize
- No: Usually don't
- Compute unit eigenvectors/values of $X^T X$
- Optional: choose k based on the eigenvalue sizes
- For the best k-dimensional subspaces, choose the k largest eigenvalues.
- Compute the coordinates $x^T v_i$ of training/test data in principle components space.

2.2 PCA Derivation 2:

Find direction w that maximized variance of projected data. Maximize

$$\text{var}(x_1, ..., x_n) = \frac{1}{n} \sum_{i=1}^{n} (x_i^T \frac{w}{\|w\|})^2 = \frac{1}{n} \frac{\|(Xw)^2\|^2}{\|w\|^2}$$

where the term on the left is the Rayleigh Quotient.

If λ_d is the biggest eigenvalue, than v_d is the vector that corresponds to that maximum variance. And if we constrain the second best w to be orthogonal to v_d, then the optimal vector is v_{d-1}

PCA derivation 3 Find direction w that minimizes projection error. Notice the similarity between this and the sum of squares.

Our optimization problem becomes the following:

$$\min_w \sum_{i=1}^{n} |x_i - \hat{x}_i|^2 = \sum_{i=1}^{n} |x_i|^2 - (x_i^T \frac{w}{|w|^2})$$

therefore maximizing projection error is equivalent to maximizing variance.

2.3 EIGENFACES

X contains n images of faces, d pixels each

- Face Recognition:Given a query face, compare it to all training fares, find nearest neighbor in $\mathbf{R}^d$.
- Each query takes O(nd) time
- Solution: Run PCA on faces to a much smaller dimension d', now neaest neighbor takes $O(nd')$ time.