Machine Learning

Unsupervised Learning

Mayson Ma

Courant Institute of Mathematical Sciences - New York University

June 2021

Table of Contents

Introduction

Idea

- We have sample points, but no labels! No classes, no y-values, nothing to predict.
- Goal: Discover structure in the data.

Examples

- Clustering: partition data into groups of 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.

Principal Components Analysis

Setting

Prior to running PCA , typically we first pre-process the data to normalize its mean and variance.

- Let X be $n \times d$ design matrix of data.
- Let mean $\mu = \frac{1}{n} \sum_{i=1}^{n} X_i^\intercal$
- Replace each $X_i^\intercal$ with $X_i^\intercal - \mu$
- Let $\sigma_j^2 = \sum_{i=1}^{n} X_{ij}^2$
- Replace each X_{ij} with X_{ij} / σ_j^2

Variance of Data

- Suppose we have a set of points $S = \{x_i\}_{i=1}^n$. On the set of S we define the uniform distribution with $\Pr(x) = 1/n$ if $x = x_i$ for some i and zero elsewhere. This probability mass function corresponds to what we call the **empirical distribution**.
- Recall that for a random vector x we have the covariance matrix

$$\Sigma = \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^\intercal]$$

The expectation is taken over the distribution of x.

Variance of Data

- The covariance matrix of a set of points is taken over this distribution which is thus defined as

$$\Sigma = \mathbb{E}[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^\intercal] = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\intercal$$

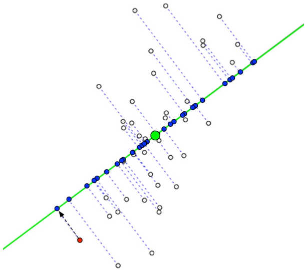
When $\bar{x} = 0$, we obtain $\Sigma = \frac{1}{n}X^\intercal X$.

- If a random vector x has covariance $\Sigma = Q\Lambda Q^\intercal$, then $z := Q^\intercal x$ has covariance Λ, and all entries of z are independent scalar random variables with z_i having variance λ_i. Since each element of z contributes λ_i randomness to the model independently from each other, $\mathrm{tr}(\Sigma) = \sum_{i=1}^{n} \lambda_i$ represents the total randomness introduced. This is the **variance** that we refer to when dealing with sets of points in $d > 1$ dimensions.

Maximum Projected Variance

Idea

- Find direction w that maximizes sample variance of projected data.
- In other words, when we project the data down, we want to keep it as spread out as possible.



Orthogonal Projection

Let w be a unit vector. The **orthogonal projection** of point x onto vector w is $\widetilde{x} = (x \cdot w)w$. If w is not unit, then $\widetilde{x} = \frac{x \cdot w}{|w|^2} w$.

Given orthonormal directions $v_1, .., v_k$, $\widetilde{x} = \sum_{i=1}^{k} (x \cdot v_i) v_i$.

Maximum Projected Variance

Optimization Problem

$$\max_w Var(\{\tilde{x}_1, ..., \tilde{x}_n\}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i \cdot w}{|w|}\right)^2 = \frac{1}{n}\frac{|Xw|^2}{|w|^2}$$
$$= \frac{1}{n}\frac{w^\intercal X^\intercal X w}{w^\intercal w} = \frac{1}{n}R(X^\intercal X, x)$$

where $R(M, x)$ is known as the **Rayleigh quotient**.

Rayleigh Quotients

- The Rayleigh quotient is defined as

$$R(M, x) = \frac{x^\intercal M x}{x^\intercal x}$$

  for a given symmetric matrix $M \in \mathbb{R}^{m \times m}$.
- The maximum value of $R(x)$ is the largest eigenvalue λ_1 of M. That maximum is achieved at the eigenvector $x = q_1$ where $Mq_1 = \lambda_1 q_1$.
- Similarly the minimum value of R equals the smallest eigenvalue λ_n of M. That minimum is attained at the "bottom eigenvector" $x = q_n$.
- If we constrain x to be orthogonal to q_1, then $x = q_2$ is optimal to achieve the maximum value λ_2.

Minimum Reconstruction Error

Idea

Find direction w the minimizes "projection error".

Optimization Problem

$$\begin{aligned}
\min_w \sum_{i=1}^{n} |x_i - \widetilde{x}_i|^2 &= \sum_{i=1}^{n} \left| x_i - \frac{x_i \cdot w}{|w|^2} w \right|^2 \\
&= \sum_{i=1}^{n} \left(|x_i|^2 - \left(\frac{x_i \cdot w}{|w|} \right)^2 \right) \\
&= \text{constant} - n \cdot Var(\{\tilde{x}_1, ..., \tilde{x}_n\})
\end{aligned}$$

Min reconstruction err or $\Leftrightarrow$ Max projection variance

Algorithm

PCA Algorithm

- Center X
- Normalize X
- Compute unit eigenvector and eigenvalues of $X^\intercal X$
- Optional: choose k based on the eigenvalue sizes
- For the best k-dimensional subspace, pick eigenvectors $v_1, \dots, v_k$
- Compute the coordinates $x \cdot v_i$ of training/test data in the principle components space

Singular Value Decomposition

Problems

- Computing $X^\intercal X$ takes $\theta(nd^2)$ time.
- $X^\intercal X$ is poorly conditioned (numerically inaccurate eigenvectors)

Fact

- Suppose $n \geq d$, we can find a singular value decomposition $X = U\Sigma V^\intercal$ where $U \in \mathbb{R}^{n\times d}, \Sigma \in \mathbb{R}^{d\times d}, V^\intercal \in \mathbb{R}^{d\times d}$.
- v_i is an eigenvector of $X^\intercal X$ with eigenvalue σ_i^2.
- We can find the k greatest singular values and corresponding vectors in $O(ndk)$ time.
- Important: Row i of $U\Sigma$ gives the principle coordinates of sample point x_i, (i.e., $x_i \cdot v_j$ for each j).

Best Low Rank Approximation

Given a matrix A, we extract its most important part A_k (**largest σ's**).

$$A_k = \sigma_1 u_1 v_1^T + \cdots + \sigma_k u_k v_k^T \qquad \text{with } \mathrm{rank}(A_k) = k$$

A_k solves a matrix optimization problem. **The closest rank k matrix to A is A_k.**

Eckart-Young

If B has rank k then

$$||\mathrm{A} - \mathrm{B}||_{\mathrm{F}} \geq ||\mathrm{A} - \mathrm{A_k}||_{\mathrm{F}}$$

The notation $||A||_F$ represents the Frobenius norm. This is equal to the square root of the sum of all the squared entries of the matrix, $||A||_F = \sqrt{\sum_{ij} A_{ij}^2}$.