Math 123
Lecture Notes

Jan 19th

Linear algebra review:

Let $A \in \mathbb{R}^{m \times n}$, with entries $a_{i,j}$, $0 \leq i \leq m, 0 \leq j \leq n$. Of course, if $m = n$, A is square.

Recall that if $x \in \mathbb{R}^n$, we say that we can act on x via $Ax \in \mathbb{R}^m$, where $y = Ax$ has entries $y_i = \Sigma_{j=1}^n a_{i,j} x_j$

Somewhat obvious remark: Vector inner products can be realized as matrix multiplication, where we realize the inner product for $x, y \in \mathbb{R}^d$ as $\Sigma_{i=1}^d x_i y_i$. Alternatively, we may consider it as matrix multiplication of xy^T, where we realize x, y as $1 \times d$ matrices.

On the other hand, if we still realize these as $1 \times d$ matrices, we can call the outer product something like $x^T y \in \mathbb{R}^{d \times d}$.

Remark: if we have a rank 1 matrix, then we have a theorem that states that we can always decompose it as the outer product of two vectors.

Recall we may define $\|x\| = \sqrt{< x, x >}$

Recall that we may call the angle of two vectors as $\arccos\left(< x, y > / \|x\|^2 \|y\|^2\right)$

Recall that, for $A \in \mathbb{R}^{m \times n}$, we call $A^T \in \mathbb{R}^{n \times m}$, where $a_{i,j} = a_{j,i}^T$.

If $A^T = A$, then we call A symmetric. Note that of course, if A is symmetric, then A must be square.

Remark: If A is symmetric, then we may decompose A as:

$$A = U^T \Lambda U$$

where

(1) U is orthogonal/orthonormal (that is, $U^T U = U U^T = I$)

(2) Λ is diagonal

In particular, the values of Λ are the eigenvalue of the corresponding eigenvector in U.

Why would we care?

1) U orthogonal $\implies \{u_i\}_{i=1}^n \subset \mathbb{R}^n$ are linearly independent, that is, they span $\mathbb{R}^n$.

2) Thus, when we want to compute Ax, we may rewrite $x = \Sigma_{i=1}^n c_i u_i$

3) Thus, when we take $y = Ax$ rewritten in the eigenbasis, this is very easy: we can compute this easily as $y = \Sigma_{j=1}^n c_j \lambda_j u_j$

Jan 24th

Principal Component Analysis:

Suppose we have data points $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$. We wish to find the directions of maximum variance in the data, and reduce the analysis into components $< D$, in order to capture as much of the variance in a lower dimensional subspace. In other words, we want to find low-dimensional structure in the data, generally when D is large.

We do this by taking a spectral value decomposition of the empirical covariance matrix.

We organize the data into a $n \times D$ matrix, where a row is a data point, and each column is a measurement of a variable across data points.

Given an orthogonal matrix $U \in \mathbb{R}^{n \times n}$, for any $x \in \mathbb{R}^n$, we may express $x = \sum_{i=1}^n c_i u_i$ where u_i is the i-th row/column of U. That is, the u_i span $\mathbb{R}^n$. However, this is not great for a generic matrix M, since this grows as the cube of the dimension. (Interesting point: solving a generic linear system grows as n^3 where n is the dimension). However, for an orthogonal matrix, we abuse the fact that $u_i \cdot u_j = \delta_{ij}$. We notice then that $Ux = (c_1, ..., c_n)$. (note that we can of course, compute each coefficient one by one by taking a dot product against each of the orthogonal vectors in turn) In particular, this has complexity n^2.

In any case, in this context, PCA is an effort to learn the "best" orthogonal matrix for my data. When we say best, we mean that if we were to take cutoffs, each choice of component preserves the maximum variance in the data, under the projection onto the component. Formally:

The projection of the data onto u_1 should be variance maximizing over all possible choices of u_1. More generally, we wish this to be variance maximizing over n components, for all n dimensional subspaces.

First, recall what variance means: for $x_1, ..., x_n \in \mathbb{R}$, we define the variance as:

$$\sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2, \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where we call μ the mean.

Then, we need only compute the projection onto a direction $u \in \mathbb{R}^{D \times 1}$ by taking $u^T x$ for a data point $x \in \mathbb{R}^{D \times 1}$, and take the variance of these pure numbers. So, we wish to find:

$$\arg \max_{u \in \mathbb{R}^n, \|u\|=1} \frac{1}{n}\sum_{i=1}^{n}[u^T x_i - \mu]^2, \mu = \frac{1}{n}\sum_{i=1}^{n} u^T x_i$$

This kinda sucks. However, we claim that WLOG, that $\mu = 0$, because we can recenter the data at the origin by taking $\overline{x} = x - \mu$, where this μ is the mean of the original data points.

Well, let's now do some tricks:

$$(u^T x_i)^2 = (u^T x_i)(u^T x_i)^T = u^T x_i x_i^T u = u^T (x_i x_i^T) u$$

where we use the fact that since $u^T x_i$ is 1×1, so $(u^T x_i)(u^T x_i)^T$

Therefore, using the linearity:

$$\frac{1}{n}\sum_{i=1}^{n}[u^T x_i]^2 = \frac{1}{n}\sum_{i=1}^{n} u^T(x_i x_i^T)u = \frac{1}{n}u^T\left[\sum_{i=1}^{n}(x_i x_i^T)\right]u$$

We call

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i x_i^T)$$

the (empirical) covariance matrix of the data, where we understand $x_i \in \mathbb{R}^{D\times 1}$, i.e. a column vector.

In this case, we can do this via Lagrange multipliers.

$$\arg\max_{u\in\mathbb{R}^n, \|u\|=1} u^T\Sigma u - \lambda\left[uu^T - 1\right]$$

Differentiating with respect to u, we end up changing this into the case where

$$\frac{\partial}{\partial u}\left[u^T\Sigma u - \lambda(u^T u - 1)\right]$$

, where we claim that this is

$$2\Sigma u - 2\lambda u = 0 \implies \Sigma u = \lambda u$$

that is, u is an eigenvector of Σ.

Jan 26th

We recall that, given data $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$, we wish to find the directions of maximum variance in the data such that $u_1 \perp u_2 \perp ... \perp u_d$, such that u_1 represents the direction of maximum variance, $\{u_1, u_2\}$ to be a plane, etc.

Recall this is an optimization problem as such:

Define $F(u) = \frac{1}{n}\sum_{i=1}^{n}(u^T x_i)^2$. Where we noticed that we could rewrite this as:

$$F(u) = \frac{1}{n}u^T\Sigma u$$

via a transpose trick.

After taking the derivatives, we recall that this becomes a eigenvalue equation. In particular, we notice that because this is symmetric, by the spectral theorem, that we have a D dimensional orthonormal basis from the eigenvectors.

Now, let's prove that the greedy algorithm works.

Suppose it works up to m vectors. Then, let's rewrite the condition for the $m+1$ vector as a Lagrange multiplier:

$$\mathcal{L}(u) = u^T\Sigma u - \sum_{i=1}^{m}\alpha_i(u^T v_i) - \alpha_{m+1}(uu^T - 1)$$

Taking a derivative again, we see that:

$$\frac{\partial}{\partial u}\mathcal{L}(u) = 2\Sigma u - \sum_{i=1}^{m} \alpha_i v_i - 2\alpha_{m+1} u = 0$$

Here, we look at this with respect to each v_i in turn.

$$v_j^T \left[2\Sigma u - \sum_{i=1}^{m} \alpha_i v_i - 2\alpha_{m+1} u \right] = 0 \implies 2v_j^T \Sigma u - \alpha_j - 2\alpha m + 1 v_j^T u = 0$$

Uh. I think James doesn't know what he's doing here.

Jan 31st

Redoing last time's proof.

Let Σ be a covariance matrix. Let $\{v_i\}_{i=1}^{D}$ be a set of orthonormal eigenvectors such that the corresponding eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_D$.

Suppose we know that the k dimensional hyperplane of greatest variance is spanned by the span of $\{v_i\}_{i=1}^{k}$. We wish to show that if we impose that u is perpendicular to v_i for $0 \leq i \leq k$, $u^T u = 1$, then $u = v_{k+1}$.

Writing this as a Lagrange multiple question:

$$\mathcal{L}(u) = u^T \Sigma u - \sum_{i=1}^{m} \alpha_i (u^T v_i) - \alpha_{m+1}(uu^T - 1)$$

Differentiating again, and setting to 0 such that this is a critical point:

$$\frac{\partial}{\partial u}\mathcal{L}(u) = 2\Sigma u - \sum_{i=1}^{m} \alpha_i v_i - 2\alpha_{m+1} u = 0$$

We claim that we may show that $\alpha_i = 0$ for $i \leq i \leq m$. If we multiply both sides by v_j^T for $1 \leq j \leq m$:

Then, since Σ is normal, and u is perpendicular to every v_j, we have that:

$$\alpha_j v_j^T v_j = \alpha_j = 0$$

Thus, we have that:

$$2\Sigma u = 2\alpha_{m+1} u \implies \Sigma u = \alpha_{m+1} u$$

Thus, u is an eigenvector, and to maximize the variance, we take an eigenvector with largest eigenvalue remaining.

What is this computational complexity of PCA?

Note that we usually just count scalar multiplications and additions and count them as equal in this toy model.

Well, suppose we have $\{x_n\}_{i=1}^{n} \subset \mathbb{R}^D$.

(1) Centering data: compute $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$. This has complexity nD

(2) Building $\Sigma = \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i$. You can do some tricks, but it's of form nD^2

(3) Eigenvectors/values of Σ: of form D^3

So, we can say that overall, this has complexity of form $D^2(n + D)$, or $D^2 \max\{n, D\}$

Note that the eigendecomposition of Σ is available here because we have a symmetric, and thus normal matrix.

However, we can also do this for a matrix $A \in \mathbb{R}^{n\times d}$, for even $n \neq d$. We will lose potentially orthogonality though between left and right eigenvectors.

The singular value decomposition of a matrix $A \in \mathbb{R}^{n\times d}$ is the factorization:

$$A = U\Lambda V^T$$

such that Λ is "diagonal", $U \in \mathbb{R}^{n\times n}$, orthogonal, and $V^T \in \mathbb{R}^{d\times d}$, orthogonal.

More precisely, Λ has only non-0 entries when, if λ_{ij} represents the i-th row, j-th column, then λ_{ij} is non-0 only if $i = j$. Hence, we denote these as λ_i.

We call the columns of U the left singular vectors of A, and the rows of V^T the right singular vectors of A, and λ_i the singular values.

Feb 2nd

Talking more about SVD, we notice that it is unique up to scaling by norm 1 scalars, and permutation of pairs of vectors.

So one way we can think of this is that we can decompose an arbitrary matrix into a sum of rank 1 matrices:

$$A = \sum_{i=1}^{\sum\{n,d\}} \lambda_i u_i v_i^T$$

with a scale factor λ_i that corresponds to its singular value.

We recall that if A is square and symmetric, then this is very nice, since the scale factors are exactly eigenvalues, and the vectors $u_i, u_i = v_i$ are exactly eigenvectors.

Recall that for any matrix, the row rank is the dimension of the rows, and the column rank is the dimension of the columns, and the overall rank is the minimum of the two.

Theorem:

The rank of $A \in \mathbb{R}^{n\times d}$ is equal to the number of non-zero singular values.

Remark, a rank 1 matrix takes at least $n + d$ pieces of data, and more generally speaking, a rank k matrix has $k(n + d)$. This is nice, because the naive approach, you may need to allocate nd pieces of data.

PCA is connected to SVD as follows:

Recall we define $\Sigma = \sum x_i x_i^T$. Define $X \in \mathbb{R}^{n\times d}$ via the rows of X being x_i as row vectors.

We notice that $X^T X = \Sigma$, so PCA wants a SVD of $X^T X$.

So, let $X = U\Lambda V^T$, its SVD. Then, expanding $X^T X$:

$$X^T X = (U\Lambda V^T)^T (U\Lambda V^T) = V\Lambda^T U^T U\Lambda V^T = V\Lambda^T \Lambda V^T$$

Here, we notice that $\tilde{\Lambda} = \Lambda^T \Lambda$ is diagonal.

So, we have that $X^T X = V\tilde{\Lambda} V^T$, which is actually its eigendecomposition.

In summary, the eigendecomposition of Σ can be written in terms of the SVD of X. So we have a correspondence between the eigenvectors of Σ and the right singular vectors of X and between the square of the singular values to the eigenvalues.

The upshot here is that sometimes, taking the SVD of the rows/columns of your data points is a similarly efficient way to get the eigendecomposition instead of hitting the covariance matrix directly, but specifically, it's numerically more stable.

Feb 7th

Unsupervised Learning: Analysis of data without human-annoatated training labels.

Clustering is one example of this. We wish to assign labels to data, without training.

More concretely, suppose we have $\{x_i\}_{i=1}^{\infty} \subset \mathbb{R}^d$. Clustering will then assign labels $\{y_j\}$ for $1 \leq j \leq K$, and under some algorithm to determine the clusters, we assign x_i to each y_j.

The most foundational clustering method is K-means, where we try to find centers from the data, and minimize the distance, somehow, over these partitions.

So now consider a functional on the space of partitions.

$$F(C_1, ..., C_k) = \sum_{j=1}^{k} D(C_j)$$

where $D(C_j)$ is a shorthand for the sum of distances of points x_j in C_i to a privileged point.

Proposition:

Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$. Define $G(x) = \frac{1}{n} \sum_{i=1}^n \|x_i - x\|^2$.

We notice that the mean minimizes G.

So here, for K-means, we choose the privileged point as the mean of a set of numbers.

So if we try to have $\mu_j = \frac{1}{|C_j|} \sum_{x_j \in C_j} x_j$, one function F can be:

$$F(C_1, ..., C_K) = \sum_{j=1}^{K} \sum_{x \in C_j} \|x - \mu_j\|^2$$

and thus, this becomes a minimization problem over the choice of partitions C_i.

But this is not easy to optimize, and it's not clear if this converges.

Scheme:

Input initial centroids $\{\mu_j\}$, with a tolerance $\epsilon = 1$.

Run an iteration. If the tolerance is more than ϵ, use centroids, assign points to the nearest centroid, and the then compute new centroids from these partitions.

If our functional has a delta from the last iteration of less than our tolerance, then we stop.

Otherwise, we continue, up to a maximum iteration count.

In general, we call this scheme a maximization expectation scheme. This specific one is Lloyd's Algorithm. This guy has an overall complexity in one iteration of order $nD + nKD$

Remark: to achieve convergence, you can potentially need $2^{\sqrt{n}}$ iterations.

Feb 9th

Recalling k-means, we notice that this is a hard assignment rule. That is, for each data point x_i, it gets exactly one label/cluster y_i. And maybe this is good, maybe this isn't ideal, and we want to include fuzziness.

Hierarchical Clustering instead gives a family of possible clusterings that relate to each other in a natural way.

Let $C = \{C_i\}_{i=1}^k$ be a clustering of data points $X = \{x_i\}$ into k clusters. That is, the C_i form a partition of X.

Let $\mathcal{B} = \{B_j\}_{j=1}^l$ be another partition of X. We call $\mathcal{B}$ nested inside of $\mathcal{C}$ if for each B_j, there exists a C_i such that $B_j \subseteq C_i$.

In particular, the family of clusterings that hierarchical clustering produces are exactly nested families. So the question here is, do you start with singletons or the whole set?

Definition:

We call a strategy that starts with singletons an agglomerative strategy. We call a strategy that starts with the full set divisive.

A remark: nesting structure is nice because we can visualize these via dendrograms, where a dendrogram is a tree that shows the splitting.

Ok, so what kind of merge rules make sense? The approach we will take is as follows:

(1) Define a distance between sets.

(2) Iteratively merge between the smallest distance sets.

So, the question is, how do we define this distance Δ? Two choices are as follows:

$$\Delta_{SL}(C_i, C_j) = \min_{x \in C_i, y \in C_j} \|x - y\|$$

which we call single linkage and

$$\Delta_{CL}(C_i, C_j) = \max_{x \in C_i, y \in C_j} \|x - y\|$$

which we call complete linkage and

$$\Delta_{GA}(C_i, C_j) = \frac{\sum_{x \in C_i, y \in C_j} \|x - y\|}{|C_i||C_j|}$$

which we call group average.

Note, a naive single linkage algorithm is $O(n^3)$, since each comparison needs around n choose 2 comparisons, so on order n^2, and we have n iterations.

Theorem [Sibson 1975]:

You can do single linkage in $O(n^2)$.

Feb 14th

In summary, we've tried clustering based off of geometric shapes (K-means) as well as distances between clusters (hierarchical clustering).

Yet another approach is to consider a notion of density.

This tends to work well when we have many outliers or a lot of noise, since we can extract out what points are noise and what are data.

Conceptually, the idea would be to cluster things that are close to their 10 nearest neighbors, or some heuristic like that.

So, a foundational method is density-based spatial clustering of applications with noise, acronym-ed "DBSCAN".

So, we want to first identify and quarantine out the ouliers, based on how many close neighbors it has. Then, we group inliers together in an iterative manner.

Concretely:

Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$.

For a point $x \in \mathbb{R}^D$, we take the closed ball $B_r(x) = \{y \in \mathbb{R}^d | d(x, y) \leq r\}$, for some metric. Unless we specify otherwise, we will assume this is the Euclidean distance.

Let $\epsilon > 0$, and define a min points between $1, n$. We take ϵ to be the threshold on how close a point needs to be to be a neighbor. We take min points to be the minimum number of neighbors within ϵ to be considered an inlier.

We call x a core point if $|B_\epsilon(x) \cap \{x_1, ..., x_n\}| \geq$ min points.

Call x a border point if x is not a core point, and there exists a core point z such that $x \in B_\epsilon(z)$.

Call x a noise point/outlier if it is neither a core nor a border point.

We then say that if x is directly density reachable from y if y is a core point, and $x \in B_\epsilon(y)$.

Further, we call x density reachable from y if there exists a sequence $\{x = x_0, x_1, ..., x_j = y\}$ such that x_{i+1} is directly density reachable from x_{i+1}.

We say x, y are density connected if there exists a core point z such that both x, y are density reachable from z.

Finally, the clusters learned by DBSCAN are maximal sets of density connected paths.

An example of when DBSCAN fails, but K-means works is a dumbbell. Everything is connected, but there are clearly two centroids. An example of the reverse is two concentric circles.

So good things: DBSCAN doesn't really care about the shapes or geometries of the situation.

Bad things, we have two parameters to tune, and it can be computationally expensive.

Remark: the notion of density is implicit in DBSCAN in terms of core, border, and noise points. We can more explicitly compute this via kernel-density estimates.

That is, let $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}$ such that:
(a) $\mathcal{K}(x) \geq 0$
(b) $\int_{\mathbb{R}^d} K(x)dx = 1$
(c) $\mathcal{K}(x) = \mathcal{K}(-x)$.
We call such a function a kernel.

The kernel density estimator of our data is:

$$\rho(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \mathcal{K}\left(\frac{x - x_i}{h}\right)$$

for some scale parameter h.

It turns out, with sufficient restrictions on n, h, that ρ converges to a nice function.