Mathematics Review of ML: COM4771

Achille Nazaret

November 12, 2019

Goal is to focus on:

- Proof writing: conciseness, rigor and clarity
- Problem solving: how to tackle a problem, translate the goal into mathematical notations, understand the manipulation, interpret and intuition
- Linear algebra: manipulation of matrices
- Euclidian spaces: Symmetric matrices, eigenvalues, spectrum

Notations:

- Kernel of a matrix is the same thing as Nullspace: $Ker(A) = Nullspace(A)$
- Set of matrices of size $n \times n$ with coefficients in field $\mathbb{K}$: $\mathcal{M}_n(\mathbb{K}) = \mathbb{K}^{n\times n}$ ($\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ in this review)

1 Warm up

1.1 Basic Algebra

We start with some algebra manipulations to warm-up. For instance, let's do some bloc matrices manipulations (it will be useful for second exercise).

Exercise 1: Determinant of bloc matrices

(a) Do you remember $\det \begin{pmatrix} A & 0 \\ B & C \end{pmatrix} = \det A \det C$? How would you prove it?

(b) Let $(a, b, c, d) \in \mathbb{R}^4$ and recall the determinant of $X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$.

Now we wish to study if this result generalizes to blocs: let $A, B, C, D \in \mathbb{R}^{n\times n}$ and study $M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$.

(c) Compute $N = M \cdot \begin{pmatrix} D & 0_n \\ -C & I_n \end{pmatrix}$

(d) Can you conclude with some additional assumptions?

(a) One on D
(b) One on B, C

(e) Prove the result without the assumption on D.

1.2 Midterm - Problem 2

Now that we are warmed-up, let's look at problem 2 from the exam.

Exercise 2: Midterm Problem 2

Let $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n) \in \mathbb{R}$ be training examples. Suppose you have a weight vector $\hat{w} \in \mathbb{R}^d$ that satisfies the normal equations for these training examples. You are now curious to see if your model is better with an affine classifier and thus want to solve the normal equations on the feature expanded training data obtained with the feature expansion $\varphi : \begin{cases} \mathbb{R}^d & \longrightarrow & \mathbb{R}^{d+1} \\ \mathbf{x} & \longmapsto & \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} \end{cases}$.
Suppose that the initial data is centered – $\sum_{i=1}^{n} x_i = 0$ – and use the previous $\hat{w}$ to compute in linear time – $O(n)$ – the new w for the affine expanded data.

Does the constant in A matter? Understand what is important. Here what is important in $\hat{w}$. It is the normal equation right? A change in the normal equations will change $\hat{w}$. If really no idea, write the hypothesis on your draft.

Exercise 3: Empirical means

Let $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n) \in \mathbb{R}$ be training examples.
Show that an optimal affine classifier (for square loss) maps the empirical mean of the features to the empirical mean of the labels.

1.3 Spectral theorem, eigenvalues and other classic tools

Some topics

- What is an eigenvalue, geometric interpretation, importance of eigenvalues with diagonalization, can all matrices be diagonalized, even complex?
- Eventually present an equivalence condition: minimal polynomial factors completely with simple roots. Sufficient condition if characteristic polynomial factorizes with simple roots as $\mu_A \mid \chi_A$
- Importance of symmetric matrices

Exercise 4: A classic result

Show that $Ker(A^T A) = Ker(A)$.

Exercise 5: Relation between A and A^t in Euclidian space

Show that $Ker(A)^{\perp} = Span(A^T)$.

Exercise 6: Eigenvalues intuition

Let V be a subspace of $\mathbb{K}^n$, we say V is an invariant subspace of A (or is stable by A) if $\forall x \in V,\ Ax \in V$.

(a) Show that every $M \in \mathcal{M}_n(\mathbb{C})$ has at least one eigenvalue

(b) Find a 2×2 real matrix that has no real eigenvalues. What is the geometric interpretation in 2d

(c) Can you find a 3×3 real matrix that has no eigenvalue? Geometric interpretation?

(d) Still in 2d, if you have one stable subspace, what can you say?

(e) Suppose $\mathbb{R}^n = B \oplus C$ and B is an invariant subspace of A. What does this implies on the bloc structure of A (if you write $A = \begin{bmatrix} D & E \\ F & G \end{bmatrix}$)

Exercise 7: Symmetric matrices

Let $A \in \mathcal{S}_n(\mathbb{R})$ be a real symmetric matrix (meaning $A^T = A$) of size $n \times n$.

1. Main property: write $\langle x, y \rangle = y^x$ the usual dot product on $\mathbb{R}^n$. Show that:
$$\forall (x, y) \in (\mathbb{R}^n)^2, \quad \langle Ax, y \rangle = \langle x, Ay \rangle$$
2. Find a counter example if A is not symmetric
This property – self adjoint operator – is actually the intrinsic condition on which spectral theory is based on.
3. Take an orthonormal basis of $\mathbb{R}^n$ $(v_1, ..., v_n)$ and consider $P = \begin{pmatrix} \uparrow & & \uparrow \\ v_1 & \dots & v_n \\ \downarrow & & \downarrow \end{pmatrix}$. Show P is inversible and $P^{-1} = P^t$
4. Following up on question (e) of last exercise, and using the first exercises we proved, can you find a nice bloc structure of the matrix A in a particular basis.
5. Prove spectral theorem by induction

- For symmetric matrices: diagonalization with **orthogonal matrices**
- Two view about diagonalization:
  1. $A = P^t diag(\lambda_i) P$
  2. $A = \sum_{i=1}^n \lambda_i v_i v_i^t = \sum_{i=1}^r \lambda_i v_i v_i^t$ with $r = rank(A)$
- The second view can be extended for rectangular matrices with the SVD (homework 3):
$$A = \sum_{j=1}^{r} \mu_j u_j v_j^T$$
- This form makes a lot of sense in regularization (follow up on homework 3)

1.4 Midterm - Problem 3

Exercise 8: Warm-up

- Show that there is a unique solution of the normal equations living in rowspace of A.
- Show that any minimal euclidian norm solution of the normal equations live in rowspace of A.

$\rightarrow$ This show the minimal euclidian norm solution is **unique** *and* **lives in rowspace** of A.

Exercise 9: Midterm - Problem 3

Let $(x_1, y_1), ..., (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$ be labeled training examples,arranged in the data matrix $A = \frac{1}{\sqrt{n}} \left[x_1 | \dots | x_n\right]^T \in \mathbb{R}^{n \times d}$ and response vector $b = \frac{1}{\sqrt{n}} \left[y_1 \dots y_n\right]^T \in \mathbb{R}^n$.
Assume $rank(A) < d$ and perform a gradient descent algorithm with step size $\eta > 0$ starting from

$$w^{(0)} = p + r, \text{ where } (p, r) \in Span(A^t) \times (Ker(A) \backslash \{0\}).$$

Is there a choice of $\eta > 0$ such that the sequence of gradient descent iterates $w^{(t)}$ converges to the minimum Euclidean norm solution to the normal equations? (Here, convergence is with respect to Euclidean distance in $\mathbb{R}^d$.)

1.5 MLE manipulation

Exercise 10: MLE with known mean but unkwown variance

- What are the MLE for $1d$ Gaussian in the following settings (MLE for the unknown parameter):
  - Unknown mean, known variance σ^2
  - Both unknown
  - Known mean μ, unknown variance

  Let $X \sim \mathcal{N}(\mu, \Sigma)$, we want to compute the MLE for Σ if we know the mean μ.
  We recall
  $$p(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} \det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^\intercal \Sigma^{-1} (\mathbf{x}-\mu)}$$
- Compute the log likelihood of $p(x_1, ..., x_n \,; \Sigma)$