Math of Data Science: Lecture 1

Vlad Kobzar

APAM, Columbia

September 6, 2022

- Course intro

Course intro

Problems involving data analysis:

- *Unsupervised learning/dimensionality reduction*
  - PCA and various other types of matrix factorization and completion
  - Problems on graphs, such as clustering
- *(Self)supervised learning*
  - regression (including sparse regression, compressed sensing, kernel methods, regularization techniques)
  - classification, including logistic regression and SVM and kernelized SVM
  - mathematical aspects of deep learning (including CNNs and models for sequential data and graphs);
- *Learning with incomplete information/policies for interaction with the environment*
  - "bandit" problems, Markov decision processes, mathematical aspects of reinforcement learning

- Combine theory and computation
  - Theory tells us about solutions and how to find them
  - Computation allows us to find solutions
  - They are related: understanding computational methods is a type of theory

Tools

- The main math tools for this course are linear algebra and probability/statistics
- The main computational tool is optimization
- Probability and statistics will come in two forms:
  - *Randomized models*: data is modeled by some unknown distribution; the problem would entail estimating that distribution
  - *Randomized algorithms*, e.g., stochastic gradient descent

Regression example

n data points $(a_1, b_1), \ldots, (a_n, b_n) \in \mathbb{R}^d \times \mathbb{R}$ organized as

- The *feature* matrix $A = \begin{bmatrix} -a_1- \\ -a_2- \\ \vdots \\ -a_n- \end{bmatrix} \in \mathbb{R}^{n \times d}$
- The *response* vector $b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \in \mathbb{R}^n$
- E.g., NOAA publishes hourly observation of temperature at various stations across the US
- Can we predict the temperature $\hat{b}_\tau$ at time τ at Yellowstone from the contemporaneous observations at other stations a_τ (e.g., if the Yellowstone sensor fails)?
  - Use the observations A and b from the periods when all the sensors were working

OLS: closed form sol'n

- Minimize the least-squares fit between the data and a linear model

$$\hat{x} = \arg\min_x R(x)$$

where

$$\begin{aligned} R(x) &= \sum_{t=1}^{n} (b_t - \langle a_t, x \rangle)^2 = \|b - Ax\|_2^2 \\ &= x^\top A^\top A x - 2b^\top A x + b^\top b \end{aligned}$$

- If A is full rank and $n \geq d$ ("big data" regime), then $A^\top A$ is positive definite
- Using the 2nd deriv test gives

$$\hat{x} = \left(A^\top A\right)^{-1} A^\top b$$

OLS: computational aspects

$$\hat{x} = \arg\min_x R(a) = \left(A^\top A\right)^{-1} A^\top b$$

- But if d is large, inverting $A^\top A$ is computationally expensive
- Use iterative optimization methods (e.g., conjugate gradient)
- Since R is convex, convergence is guaranteed; can study rates

OLS: stats interpretation

- If
$$b \sim N(Ax, \sigma^2 I) = Ax + N(0, \sigma^2 I)$$
OLS is the value of x that makes the data most probable, i.e.
$$\hat{x} = \arg\min_x R(a) = \arg\max_x L(x, \sigma^2) = x_{MLE}$$
where
$$R(x) = \|b - Ax\|_2^2$$
- Maximize the log of the likelihood fcn L w.r.t. x and σ^2:
$$L(x, \sigma^2) = p(b|Ax, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\|b-Ax\|^2}{2\sigma^2}}$$
- Again use the 2nd deriv test

OLS: geometric interpretation

$$\hat{x} = \arg\min_{x} R(x)$$

where

$$R(x) = \|b - Ax\|_2^2$$

- $A\hat{x} = UU^T b$ - projection of b on the span of the columns of A
- Prove using the SVD: $A = U\Sigma V^T$.

Overfitting

- Small error on the training set, but high error on a test set because $\hat{x}$ will fit the features that may not be relevant (e.g., sensors very far from Yellowstone)
- Can we find a sparse linear model?
  - E.g., predict the Yellowstone temperatures based on observations from a small subset of the stations
  - This subset is “learned” from the training data

Sparse regression (LASSO)

- l_0 penalty: $\|x\|_0 = \#$ of nonzero entries of x
- This regularization enforces sparsity: for $\lambda > 0$

$$x_0 = \arg\min_x \Big(R(x) + \lambda \|x\|_0 \Big)$$

- But is intractable (the objective not convex; l_0 not a norm)
- Would a "relaxation" to the l_1 norm also promote sparsity?
- LASSO
  - *Penalized form*

$$x_1 = \arg\min_x \Big(R(x) + \lambda \|x\|_1 \Big)$$

  - *Equivalent to constrained form*

$$x_1 = \arg\min_{\|x\|_1 \leq r} R(x)$$

  - Pf. by a Lagrange multiplier-type calculation

LASSO

- Constrained form

$$x_1 = \arg\min_{\|x\|_1 \leq r} R(x)$$

- By completing the squares,

$$R(x) = (x - \hat{w})^\top AA^T(x - \hat{w}) + R(\hat{w})$$

where the OLS solution $\hat{w}$ of the unconstrained problem is the center of the ellipsoid OLS level sets

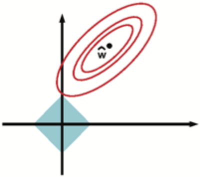

Figure: Level sets of $R(x)$ in red and the area satisfying $\|x\|_1 \leq r$ in blue (Fig 13.3 from [2]).

Solving LASSO numerically

- No general closed form solution
  - Even for OLS, the closed form solution is not used for large data sets due to computation cost of matrix inversion
- Since LASSO can be reduced to a convex optimization problem (QP), can use standard iterative solvers
- Can be more efficient to use other methods that exploit the structure of the lasso objective, e.g., the linear separability of the l^1 norm

Sparse inverse problems

- If b in column space of A and $n < d$ ("inverse problem" regime, e.g. MRI), then $Ax = b$ is an underdetermined system.
- But with sparsity and other technical assumptions, l_1 minimization can exactly recover a sparse vector x.
- (Candes, Tao, Donoho) For

$$x^* = \arg\min \|x\|_1$$
$$s.t. \; b = Ax$$

  if the row of A are not too localized so that they won't miss the entries of S-sparse x and if there is enough data $n \geq O(S \log d)$
- key idea entails recovering the support of x (i.e., indices of nonzero entries) and therefore reducing it to a well-posed problem.

Matrix completion

- Low rank models are common when only a few factors explain the variance in data organized in the matrix.
- Motivation: Netflix competition

$$\begin{array}{l} \\ \text{The Dark Knight} \\ \text{Spiderman 3} \\ \text{Love Actually} \\ \text{Bridget Jones's Diary} \\ \text{Pretty Woman} \\ \text{Superman 2} \end{array} \begin{array}{c} \begin{array}{cccc} \textit{Bob} & \textit{Molly} & \textit{Mary} & \textit{Larry} \end{array} \\ \begin{pmatrix} -10 & -10 & 10 & 5 \\ -7 & -10 & 8 & 10 \\ 8 & 10 & -5 & -9 \\ 10 & 4 & -6 & -10 \\ 8 & 9 & -9 & -4 \\ -9 & -8 & 9 & 10 \end{pmatrix} \end{array} := A,$$

- To make a recommendation, estimate missing entries

$$\begin{array}{lcccc} & \textit{Bob} & \textit{Molly} & \textit{Mary} & \textit{Larry} \\ \text{X-Men 7: Mutant Mosquito} & (-10 & ? & 8 & 10\ ) \end{array}$$

- Fit a low rank model using the SVD: $A = U\Sigma V^T$
  - a truncated rank-k SVD is the best rank-k approximation of A

Matrix completion

- Low rank structure implies correlation between entries
- *Netflix problem*: How do we exploit it to predict missing entries?
- E.g. where a user is going to like a new movie
- E.g., if the below matrix is rank 1, then we must have 1 in place of the missing entry.

$$\begin{pmatrix} 1 & ? & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \mathbb{1}\mathbb{1}^T$$

- This seems like an easy matrix to complete.

Matrix completion

- On the other hand, if a matrix is sparse or its rows correlate with the canonical basis, it seems much harder to complete

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & ? \end{pmatrix} = \begin{bmatrix} 0 \\ 0 \\ ? \end{bmatrix} \begin{bmatrix} 0 & 0 & ? \end{bmatrix}$$

  Therefore differences in the structure of a low rank matrix may determine how hard or difficult it is to complete.
- Coherence (or localization of rows and columns) introduced previously is relevant here: for $A = U\Sigma V^T$
- For example, if the left singular vectors (columns of U) correlate with the canonical basis vectors, matrix will hard to complete.

Nuclear norm minimization

- Since rank is not a convex function, minimization of the rank subject to known entries $A_0 = \{(i,j), a_{ij}\}$ is not computationally tractable.

$$
\begin{aligned}
(N) \qquad & \min \operatorname{rank}(A) \\
& A \in \mathbb{R}^{m \times n} \\
& A_{ij} = a_{ij} \text{ for } (i,j) \in A_0
\end{aligned}
$$

- Note that rank is an l_0 "norm" of Σ for $A = U \Sigma V^T$.

Nuclear norm minimization

- Instead use a "convex relaxation" based on minimization of the nuclear norm:

$$
\begin{array}{ll}
(N) & \min \|A\|_N \\
 & A \in \mathbb{R}^{m \times n} \\
 & A_{ij} = a_{ij} \text{ for } (i,j) \in A_0
\end{array}
$$

  where

$$\|A\|_N = \sum_{i=1}^{r} \sigma_i$$

  and σ_i are singular values and r is rank of A
- Note that rank is an l_1 "norm" of Σ.

Movie ratings - policies for interacting with the environment

- Let a feature vector x describe a user
- We choose 1 out of 5 hit movies and recommend it to x
- We only get the feedback on the recommended movie
- Let's say the feedback is 3 out of 5 stars
- Next time we have a similar user $x' \approx x$, should we recommend the same movie?
- Or try a different one hoping to get 5 stars?

k-armed bandit

In each $t \in [T]$,

- Environment samples reward $(X_t, R_t) \in \mathcal{X} \times \mathbb{R}^k$ from a fixed k-dimensional distribution P i.i.d.
- X_t is revealed to the player
- The player selects $A_t \in [k]$ based on history

$$\mathcal{D}_t = (A_{1:t-1}, R_{1:t-1}, X_{1:t})$$

- Player receives the reward $R_t(A_t)$

- $R_t(a)$ for $a \neq A_t$ ("counterfactuals") are not revealed to the player
- A_t is not independent from R_t - information about R_t can propagate to A_t through X_t
- But A_t is conditionally independent from R_t given X_t - R_t is not revealed to the player when it selects A_t.

Optimal policy

- Suppose we knew

$$r(x, a) = \mathbb{E}[R|A = a, X = x]$$

which gives the expected reward for each action.

- Then the optimal policy would be

$$\pi_t^*(a|X_t, \mathcal{D}_t) = \mathbb{1}[a = \arg_a \max r(X_t, a)]$$

- Here choosing an action according to policy π_t means choosing A_t randomly s.t.

$$P(A_t = a) = \pi_t(a|X_t, \mathcal{D}_t)$$

- Of course we don't know $r(x, a)$, but can we estimate it?

Next steps

- Review of linear algebra, probability and optimization
- PCA, least squares

References I

[1] Carlos Fernandez-Granda, *DS-GA 1013 / MATH-GA 2821 Mathematical Tools for Data Science, Lecture Notes*, 2020

[2] Kevin P. Murphy, Machine Learning: a Probabilistic Perspective, MIT Press, 2012

[3] David Rosenberg, *DS-GA 1003 Machine Learning and Computational Statistics, Lecture Notes*, 2017

[4] David Rosenberg, *DS-GA 3001: Tools and Techniques for Machine Learning, Lecture Notes*, NYU Fall 2021, `https://github.com/davidrosenberg/ttml2021fall`

[5] Hastie, Tibshirani, Wainwright, Statistical Learning with Sparsity: The Lasso and Generalizations, Chapman & Hall/CRC Monographs on Statistics and Applied Probability, 2015