

Past Papers unofficial solutions

While studying and unable to have the solutions for these papers, I decided to create this document which will contain all (unofficial) solutions from every question from every past paper in the MLPR course. There are only a few days till the exam, so I won't be writing answers in details; rather a sentence or a phrase to which you can use to expand your solution for that question.

These answers have been written by me or with some assistance from someone who knew the topic or material more than I do. Some questions contain multiple possible answers. It's important to note that **these answers are not official and not fact checked with a course organizer**. Take them with a grant of salt.

If you have any suggestion or any edits I should make, please email me and I will immediately act on it.

2021 Dec

1. A) (*Pick one solution*)
 - i. Gaussian Elimination -> epoch form -> # of Non 0 rows $\Sigma(2)$, so not full rank $\therefore \Sigma(2)$ causes the problem.
 - ii. Cholesky Decomposition – `numpy.linalg.cholesky(Σ)` -> $\Sigma(2)$ throws an error.
2. A) (*Any two solutions*)
 - i. **(possibly wrong)** L2 Regularization to the diagonal of the matrix – makes it positive definite.
 - ii. **(possibly wrong)** Use Partial Eigenvalue Decomposition on $\Sigma(2) = Q\Lambda Q^T$. replace negative eigenvalues with 0 -> reconstruct $\Sigma(2)$.
 - iii. Modify training data.
 - iv. Remove 3rd feature.

B)

 - i. **(mod required)** Function based on description (large values close to edges): sigmoid with range (3,5): $a_{t+1} = 3 + \sigma(\eta \cdot \nabla_a c(a_t)) \cdot 2$, with cost function $c(a) = \frac{1}{2m} \sum_{i=1}^m (y_i - f(x_i; a))^2$ where the partial derivative is: $\frac{\partial c}{\partial a} = -\frac{1}{m} \sum_{i=1}^m (y_i - f(x_i; a)) \frac{\partial f(x_i; a)}{\partial a}$. Use a small learning rate, $\eta = 0.001$.
 - ii. ☹
 - iii. Apply Gaussian Naives optimization and then RMSE (check Assignment, Q5).
3. A)
 - i. D features, N data points -> $O(D \cdot N)$
 - ii. Full Covariance -> $O(N \cdot D^2)$

B)

- i. Pro: Faster computationally.
Pro: Suitable for large databases, thanks to the use of mini batches.
Con: Harder to implement.
Con: Higher Computational Complexity (greater Big "Oh").

4. A)

- i. W: H x D (H hidden neurons, D dimensional layer x) order matters!
- ii. b: H x 1

B) Matrix A: H x K

Matrix B: K x D

Bias c : H

Extra parameters per country: H x K + K x D + H.

C) $X @ (W + A @ B) . T + c$ [None, :]

D) Less risk on overfitting, less likely to memorize training data, less memory consumption, more generative. (pick any two of these reasons)

Handwritten mathematical derivations for backpropagation through matrix multiplication:

- $h = g((W + AB)x + c)$, using $(=g(A) \Rightarrow \bar{A} = g'(A) \odot \bar{C})$
- $D = (W + AB)x + c \Rightarrow h = g(D) \Rightarrow \bar{D} = g'(D) \odot \bar{h}$
- $E = (W + AB)x$ using $C = A + B \Rightarrow \bar{C} = \bar{A}$ and $\bar{C} = \bar{B}$
- $D = E + c \Rightarrow \bar{D} = \bar{E}$ and $\bar{D} = \bar{c}$
- $F = W + AB$, $G = AB$ using $C = AB \Rightarrow \bar{A} = \bar{C} B^T$, $\bar{B} = A^T \bar{C}$
- $E = Fx \Rightarrow \bar{E} = \bar{F} x^T$
- $F = W + G \Rightarrow \bar{F} = \bar{W} + \bar{G}$
- $G = AB \Rightarrow \bar{A} = \bar{G} B^T$, $\bar{B} = A^T \bar{G}$
- Propagating: $\bar{B} = A^T g'((W + AB)x + c) \odot \bar{h} \cdot x^T$

E)

F) Initialize AB to be a zero matrix by setting the matrix B to zero. Replace bias c to b. For the first epoch, the model will be back to its original form. The gradients from B will slowly be generated after training.

G) Early Stopping (to prevent overfitting).

5. A) ☹

B) (possible idea) take $k(x_1, x_2) = E[f_1 * f_2] - E[f_1] * E[f_2]$, where $f = a + bx + cx^2$. No, we can't have a gaussian process $k(x_1, x_2) < 0$.

2021 Dec Comments

Harder than expected. The exam was open book, so questions are expected to be harder and require higher levels of technical knowledge to answer them. Don't let that bring you down. ☺
Keep in mind, that the above answers are **NOT** fact checked by any organizers of this course, so mistakes are expected to exist. The answers will be updated and checked daily by myself and my peers until the day of the examination.

2019 Dec

- i. A)
 $\mathbb{E}_{p(\mathbf{x}, y)}[L(y, f(\mathbf{x}))]$ Where: L = loss, y = training result, $f(\mathbf{x})$ = model function, p = true data distribution.
OR $\int L(y, f(\mathbf{x}))p(\mathbf{x}, y)d\mathbf{x}dy$ Where: L = loss, y = training result, $f(\mathbf{x})$ = model function, p = true data distribution.
OR The expected loss that the model (function) will achieve/yield/score on future/unseen test datapoints.
 - ii. $\frac{1}{M} \sum_{m=1}^M L(y^{(m)}, f(\mathbf{x}^{(m)}))$
 - iii. $\sqrt{\frac{1}{M} \sum_{m=1}^M (L_m - L)^2}$
 - B)
 - i. —
 - ii. —
 - C)
 - i. Generalization error doesn't predict risk to overfitting. Model could be prone to overfitting.
 - ii. Provide validation data sets to test the model and check if it overfits.
2. A) $W: K \times D$, $c: D \times 1$, $v: 1 \times K$, $b: 1 \times 1$.
B) Gradients will be zero and the model will not learn.
C) Prevents from the gradients (derivatives) from growing too large, causing optimization issues.
- 3.

2019 Dec Comments

1)a) has been fact checked from the last exam preparation lecture.