Recitation 9: Project

CIS 5450

Motivation for The Project

Studying machine learning != doing machine learning
(both are important!)

- Data cleaning, entity linking
- Creating appropriate schema(s) for table(s)
- Dealing with vast quantities of data using Spark/SQL
- Creating appropriate visualizations
- Figuring out which model is appropriate
- Evaluating models with appropriate metrics (beyond accuracy)
- Model tuning

Basic (batch) gradient descent. Gradient descent starts with some initial parameter values $\mathbf{W}$,[3] and then on each iteration, updates each parameter $w_{jk}^{(l)}$ as follows:

$$\begin{aligned} w_{jk}^{(l)} &\leftarrow w_{jk}^{(l)} - \eta \frac{\partial J(\mathbf{W})}{\partial w_{jk}^{(l)}} \\ &= w_{jk}^{(l)} - \eta \frac{1}{m} \sum_{i=1}^{m} \frac{\partial J_i(\mathbf{W})}{\partial w_{jk}^{(l)}}, \end{aligned}$$

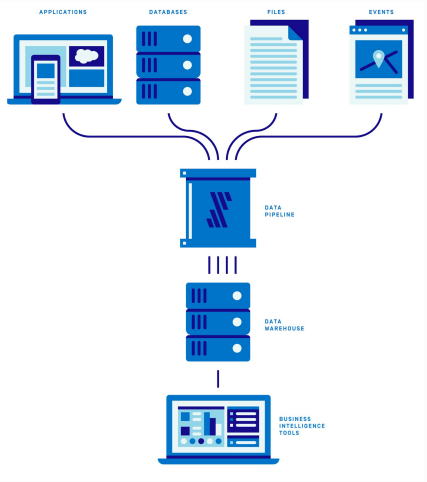
where $\eta > 0$ is the step size or learning rate parameter.[4] This process is repeated until the parameter estimates converge (or for some suitably large number of iterations).

In order to implement gradient descent, we need to compute the derivatives $\partial J_i(\mathbf{W})/\partial w_{jk}^{(l)}$. This is where backpropagation comes in; it is essentially an application of the chain rule of differentiation. Specifically, continuing with the one-hidden-layer example above, for derivatives w.r.t. weights $\mathbf{W}^{(2)}$, we have for each $j \in \{1, \ldots, d_1\}$:

$$\begin{aligned} \frac{\partial J_i(\mathbf{W})}{\partial w_{1j}^{(2)}} &= \frac{\partial J_i(\mathbf{W})}{\partial f_{\mathbf{W}}(\mathbf{x}_i)} \cdot \frac{\partial f_{\mathbf{W}}(\mathbf{x}_i)}{\partial w_{1j}^{(2)}} \\ &= 2\big(f_{\mathbf{W}}(\mathbf{x}_i) - y_i\big) \cdot a_{ij}^{(1)}, \end{aligned}$$

where $a_{ij}^{(1)} = g(z_{ij}^{(1)}) = g({\mathbf{w}_j^{(1)}}^{\top} \mathbf{x}_i)$ is the feature/activation value of the j-th hidden unit on example i (under the current weights). Similarly, for derivatives w.r.t. weights $\mathbf{W}^{(1)}$, we have for each $j \in \{1, \ldots, d_1\}$ and $k \in \{1, \ldots, d\}$:

$$\begin{aligned} \frac{\partial J_i(\mathbf{W})}{\partial w_{jk}^{(1)}} &= \frac{\partial J_i(\mathbf{W})}{\partial f_{\mathbf{W}}(\mathbf{x}_i)} \cdot \frac{\partial f_{\mathbf{W}}(\mathbf{x}_i)}{\partial w_{jk}^{(1)}} \\ &= \frac{\partial J_i(\mathbf{W})}{\partial f_{\mathbf{W}}(\mathbf{x}_i)} \cdot \frac{\partial f_{\mathbf{W}}(\mathbf{x}_i)}{\partial a_{ij}^{(1)}} \cdot \frac{\partial a_{ij}^{(1)}}{\partial z_{ij}^{(1)}} \cdot \frac{\partial z_{ij}^{(1)}}{\partial w_{jk}^{(1)}} \\ &= 2\big(f_{\mathbf{W}}(\mathbf{x}_i) - y_i\big) \cdot w_{ij}^{(2)} \cdot g'(z_{ij}^{(1)}) \cdot x_{ik}. \end{aligned}$$



Rubric Walkthrough

See Project Description document (link on Ed)

Criterion A: Project Proposal / Intermediate Check-In (5 points)

Essentially free extra-credit if you show quality progress and meet the deadlines.

Proposal: a brief high-level description of what you plan to do and what are your plans for the project (already done at this point)

Intermediate Check-In: Zoom meeting with your assigned TA during Week of **04/10 - 4/16**

Criterion B: Difficulty (10 points)

Sophistication and time commitment required for this project.

- Have you attempted ***challenging analysis?***
- Did you avoid choosing a topic that would just state the obvious and end up following the path towards confirmation bias?
- Have you thought critically about how to approach this topic?
- How much time would have been required to complete your project?

Criterion C: Code Quality / Readability (10 points)

Is your project notebook understandable and fairly well broken into modular steps?

- Code blocks are broken out logically
- Comments are present in every code block
- Every non-trivial section of a code block is clearly explained
- Easy for a third person to look at your code and tell the story of what's going on

Criterion D: Creativity / Uniqueness (10 points)

Does your project stand out from the rest?

- What makes this project unique relative to similar topics?
- Did you do EDA beyond “standard recipe” of checking for dataset size?
- Did you attempt to create a diverse set of informative plots and ask *relevant* business questions specific to the case?
- Effort awarded for joining extra datasets together or creating your own dataset

Criterion E: Visualization (20 points)

Visualize your findings well using plots and graphs.

- Ensure that they are informative, appealing, and professional
- Always think from the perspective of a stakeholder/client:
  - Is everything self explanatory and clear from the chart?
- Attention to detail really matters!
- Exceptional projects use packages such as **tensorboard**, **plotly**, **seaborn**, on top of the traditional matplotlib to create stunning and informative visualizations
- There will be a **HUGE** emphasis on accurate applications of visualizations
  - Ex: if you want to plot prices over time, you should be using a line chart over a bar chart

Criterion F: Modeling (20 points)

Supervised and Unsupervised Learning Models

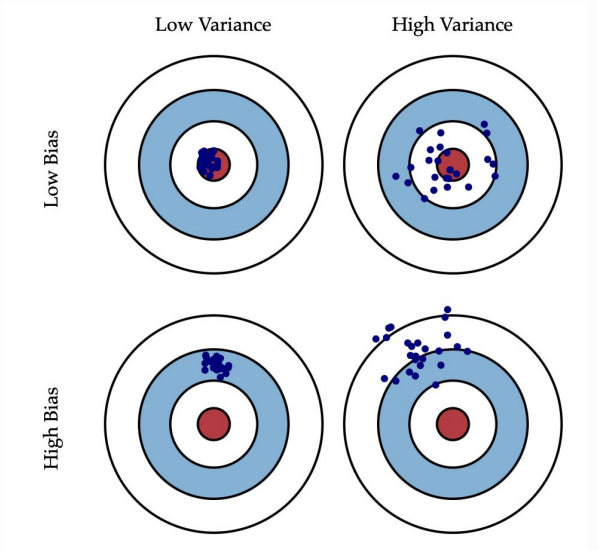
- Is your model useful and implemented correctly?
- Did you justify your choice of models that you chose to implement?
- You **must have at least 2 different models**
- You **must provide scoring metrics for each model** in order to compare the models in the end and determine the best one
    - Note: There is no minimum benchmark performance we expect from your models but just ensure you justify why a model may potentially be not doing well

Evaluating Models

"All models are wrong, but some are useful"
- George Box, statistician (1919 - 2013)

Think beyond accuracy when evaluating Classification models!

- Confusion matrices
- Precision, Recall, F_1 scores
- ROC curves, AUC-ROC score
- Bias-Variance tradeoff
    - Underfitting vs. Overfitting



Criterion G: Application of Course Topics (20 points)

Strictly evaluated against Modules 1 - Module 26

- Is your project built around the topics discussed in class, recitation, or covered on the homework assignments?
- You are welcome to go beyond the scope of the course, but you should apply a significant portion of this course's topics.
- Topics include:
  - Joining >=2 tables together
  - Usage of Pandas SQL, Pandas or SparkSQL
  - Unsupervised with Supervised Learning Techniques

Note: *going beyond the scope of class doesn't earn any credit in this criterion (it gets rewarded elsewhere, but note that the upside is capped)*

Criterion H: Quality of Final Deliverable (10 points)

Presentation matters!

- Is your final product clean, polished, and professional?
- Have you created a deliverable that is engaging and informative?
- Would this be the work you submit to a Hiring Manager for a take-home project?

Markdown Cheatsheet

Deliverable Options

1) Annotated Notebook

2) Medium Blog Post

3) Live Presentation

Sample Project Walkthrough

Big Wins:

- Extremely thorough EDA
  - Entity linking, analysis of multicollinearity
- Ample English analysis / descriptions
- Interactive data visualizations (created with Plotly)
- Every code cell is commented
- Notebook is clearly organized