SONAT

Machine learning quiz (Solution)

Created by: Tommy Odland

Instructions

Please read the entire problem sheet before starting.
If you get stuck, move to a different problem.

1. **Multiple choice**

   (a) Linear models are linear. But with respect to what exactly?

   ◯ Input variables x_i ◯ Features ϕ_i ◯ The function space

   Solution: Features ϕ_i.

   (b) Which loss function in a linear model deals well with outliers?

   ◯ ℓ_1 loss (manhattan) ◯ ℓ_2 loss (euclidean) ◯ ℓ_∞ loss (minmax)

   Solution: ℓ_1 loss (manhattan).

   (c) What concept(s) do we associate with logistic regression?

   ☐ AI ☐ GLM ☐ Binary classification

   Solution: Generalized Linear Model and Binary classification. Those who are very active on LinkedIn may include AI too.

   (d) Which abstract class(es) are in the `sklearn` API?

   ☐ Predictor ☐ Regressor ☐ Transformer

   Solution: Regressor is not a class, but the other two are.

   (e) Which two of these are sensible machine learning use cases?

   ☐ Deciding if sending emails leads to more sales
   ☐ Predicting student grades in absence of exams due to coronavirus
   ☐ Predicting which horse to bet money on in horse races
   ☐ Deciding how to plan infrastructure in a city
   ☐ Sorting an array
   ☐ Finding the best way to transmit signals through a large network
   ☐ Modeling the spread of disease
   ☐ Predicting incoming calls in a call center

   Solution: Horse races and call center are great use cases. The email problem has a more statistical flavor. The rest are not pure ML use cases, though prediction might be part of the task.

(f) What is the primary advantage of LASSO (ℓ_1 regularization)? ◯ Rotational invariance ◯ Smaller weights ◯ Sparse weights

Solution: Sparse weights.

(g) How do we deal with nominal data (e.g. colors) in a linear model?
◯ Remove it ◯ One-hot-encoding ◯ Use imputation

Solution: One-hot-encoding.

2. **Write a sentence answering these questions**

(a) What are the advantages and disadvantages of linear models compared to more complex models? (for instance neural networks)

Solution: Advantages: Train faster, predict faster, always converge to optimal solutions, easier to understand and interpret, easier to deploy, act as baseline when comparing with other models. Disadvantages: Must do more feature engineering, not competitive for complex tasks (lots of structure, lots of data, lots of non-linearity).

(b) What are common ways of dealing with missing data entries?

Solution: Remove the rows completely, use simple imputation (e.g. use mean of variable) or use a model for imputation.

(c) What is regularization, and why is it often a good idea?

Solution: To control model complexity, overly complex models tend to overfit the training data. Overfitting causes the model to not generalize well to unseen data.

(d) What is boosting?

Solution: Putting many models in sequence, where model i learns from the errors of model $i-1$. The most common model to use is trees.

(e) What is the purpose of the test data set? How is it used?

Solution: To assess how well the model will generalize to unseen data. Put the test data way, and never look at it until you are finished. Only use it to assess generalization of the final model.

(f) Name three things to think about before putting ML in production.

Solution: Concept drift. Re-training. Logging and alerts. Interpretation by business. Model history. Privacy concerns. Data quality.