Guide to Writing a Machine Learning Methods Section

Peter Sadowski

April 11, 2025

This is a guide for describing machine learning (ML) projects. It should be used when submitting project reports in my ML classes and when writing a Methods section of a research paper.

Step-by-Step

1. **Introduce the problem.**
   - What problem are you trying to solve? What will the model be used for?
   - What are the inputs and outputs of your model? ML models approximate functions; what function are you approximating?
   - Why do you expect ML to be appropriate task? How much training data do you have? Why do other methods fail?
2. **Introduce the dataset.** ML is about learning from data and the model will only be as good as the dataset used to train it. Clearly explain where your data came from and how many samples you have. The number of samples, along with the inputs and outputs, are key details that should almost always be in the abstract.
3. **Introduce the model.** Which machine learning model(s) are you using? Is there any particular reason? Different models have different **inductive biases**, and choosing the "correct" model can make a big difference in performance.
4. **Specify features and pre-processing.** Describe the feature representation and any transformations applied. This may include:
   - Feature units, e.g. inches, meters, or unit-less.
   - Feature properties, e.g. categorical, ordinal, integer, real-valued, one-hot.
   - Transformations, e.g. min-max scaling, standardization, $\log(x)$ or $\log(x+1)$.
   - Missing data methods, e.g. removal, padding, interpolation, imputation, iterative imputation.
   - Augmentation, e.g. Gaussian noise.
5. **Specify how the data is used to evaluate the model.** Cross-validation is the primary method of evaluating machine learning models. This usually consist of splitting the data into train, valid, and test datasets, or can involve more complex methods such as K-fold cross-validation, or nested K-fold cross-validation. It is important to clearly communicate how you use the dataset in order to convince the reader that you appreciate the dangers of overfitting.

   Example: *The total dataset contained 1,000,000 examples. The data was randomly permuted, then divided into three subsets: 60% training, 20% validation, and 20% test. Models were trained on the training set, while the validation set was used for early stopping, hyperparameter optimization, and model selection. The test set was used to evaluate the final model.*

6. **Specify the hyperparameter search space.** This is the list of hyperparameters that were optimized, and the range of values that were explored. This can be difficult to describe succinctly since hyperparameter tuning is usually an iterative process involving the experimenter. Ideally, you use a hyperparameter optimization framework like OPTUNA, but in the case where hyperparameter optimization was mostly done by hand, you can simply state the range of values you explored (min and max) for each hyperparameter and the total number of models you tried.

   Example: *For the K-Nearest Neighbor classifier we tried different values of K and different distance metrics. We tried all odd values of K from the set of integers between 1 and 99, {1,3,5,...,99}. We tried the L1 and L2 distance metrics.*

7. **Explain how hyperparameters were optimized.** Explain if you tuned the hyperparameters by hand, or exhaustively tested every hyperparameter combination in the search space. This can be a simple statement of the metric and validation set used. State any optimization algorithms you used, e.g. Random Search, Grid Search, Bayesian Optimization, Population-Based Training. Cite any software packages you used to implement these algorithms (e.g. OPTUNA or SHERPA).

   Examples:

   (a) *After trying all combinations of hyperparameters in the search space, the model with the highest accuracy on the validation set was selected.*

   (b) *A total of 50 different models were trained, with random combinations of hyperparameters selected from the search space. The model with the highest validation set MSE was selected and evaluated on the test set.*

   (c) *Hand-tuning was used to train a total of 20 different models with different hyperparameters. The model with the best validation set AUROC was selected and evaluated on the test set.*

8. **Evaluate the model on clean test set.** Remember that a clean evaluation requires a held-out test set that was never used to make any decisions about the data. This includes model selection, hyperparameter choices, early stopping, and data pre-processing. In practice, this can be difficult, so it is OK to make exceptions to this rule if you know the overfitting risks are small. However, in your methods section you need to make it very clear that you are aware of the risks and have taken care to get a clean evaluation; if you don't, the reader won't trust your results.

   When quantifying performance, remember to specify the *metric* and *dataset* for every number you present.

   Example:

   (a) *The performance of the final model on the test set was 0.98 AUROC.*

   (b) *The model achieved an accuracy of 80% on the held-out test set.*

9. **Comment on the risk of Distribution Shift.** There are many types of distribution shift that can cause the performance to degrade when If possible, provide justification for why you expect the model to generalize despite differences.

   Example:

   (a) *The model was trained and evaluated on a single data sample through cross-validation. Generalization performance on new data will depend on whether that data comes from the same distribution.*

   (b) *The model will be used to make predictions on data from a later time, so temporal data drift could harm model performance.*