Considerações Práticas

INFORMAÇÃO,

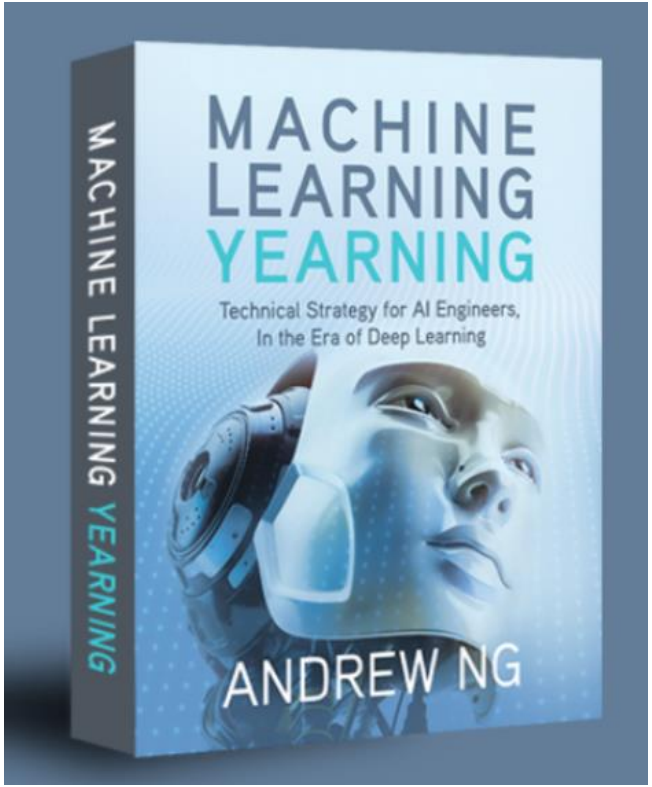
TECNOLOGIA

& INOVAÇÃO

Como escolher qual método usar?

- Considerações “estatísticas”: tamanho amostral, número de features, estrutura esperada
- Necessidade de interpretação
- Facilidade de implementar a solução final
- Custo computacional (armazenamento e tempo) para calcular as predições

https://github.com/ajaymache/machine-learning-yearning

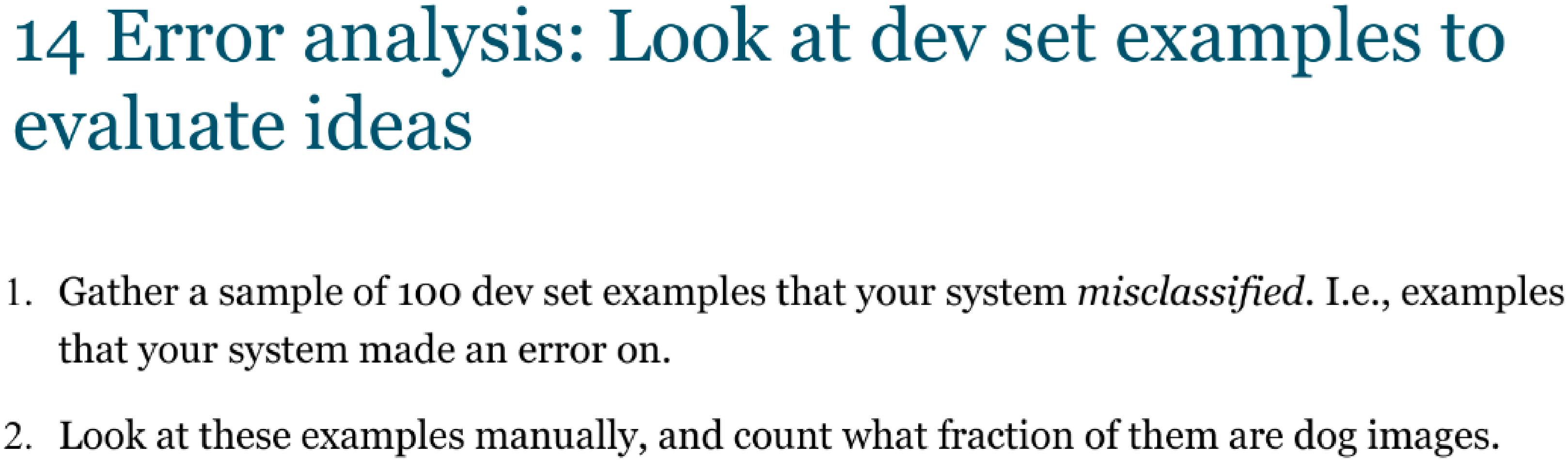

MACHINE LEARNING YEARNING
Technical Strategy for AI Engineers, In the Era of Deep Learning
ANDREW NG
MACHINE LEARNING YEARNING

> Choose dev and test sets to reflect data you expect to get in the future and want to do well on.

In other words, your test set should not simply be 30% of the available data, especially if you expect your future data (mobile phone images) to be different in nature from your training set (website images).

How about the size of the test set? It should be large enough to give high confidence in the overall performance of your system. One popular heuristic had been to use 30% of your data for your test set. This works well when you have a modest number of examples—say 100 to 10,000 examples. But in the era of big data where we now have machine learning problems with sometimes more than a billion examples, the fraction of data allocated to dev/test sets has been shrinking, even as the absolute number of examples in the dev/test sets has been growing. There is no need to have excessively large dev/test sets beyond what is needed to evaluate the performance of your algorithms.

If you are trading off N different criteria, such as binary file size of the model (which is important for mobile apps, since users don’t want to download large apps), running time, and accuracy, you might consider setting N-1 of the criteria as “satisficing” metrics. I.e., you simply require that they meet a certain value. Then define the final one as the “optimizing” metric. For example, set a threshold for what is acceptable for binary file size and running time, and try to optimize accuracy given those constraints.



14 Error analysis: Look at dev set examples to evaluate ideas

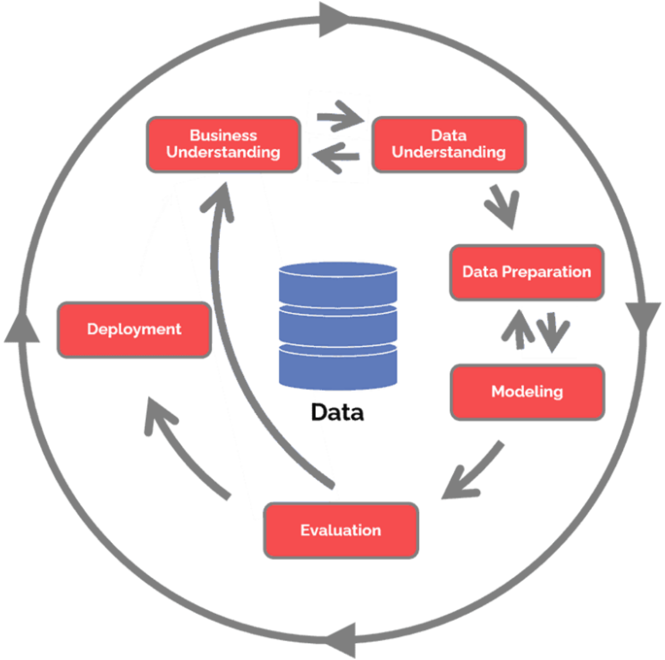
1. Gather a sample of 100 dev set examples that your system *misclassified*. I.e., examples that your system made an error on.

2. Look at these examples manually, and count what fraction of them are dog images.

If your error rate on the training set is 15% (or 85% accuracy), but your target is 5% error (95% accuracy), then the first problem to solve is to improve your algorithm's performance on your training set. Your dev/test set performance is usually worse than your training set performance. So if you are getting 85% accuracy on the examples your algorithm has seen, there's no way you're getting 95% accuracy on examples your algorithm hasn't even seen.

Viés de Seleção

Projetos de ciência de dados



Próximos passos

- A cada dia, surgem novas técnicas e algoritmos: estudo constante!
- Há um mundo de possibilidades: defina bem seus objetivos profissionais e se especialize de acordo com eles!
- Entenda bem o problema, dados e outros recursos
- Foque em resultados
- Defina processos
- Seja multidisciplinar e proativo