A quick summary of several papers within the computer science and psychology literature to answer the following questions:

1. How do we measure the accuracy of the prediction (what objective function is to be used).
2. How good does the prediction have to be before it's a test?

1 Objective Functions

The papers here which predict personality traits fall into two primary categories: predicting personality from samples of written language, and on the other side, predicting personality from usage of devices or networks.

Two papers on language both use Pearson Correlations, Arnoux et al. 2017 and Schwartz et al. 2013, which they use with an explicit explanation that personality is a continuous value, and report correlation scores between 0.25 - 0.42 on large sample sizes (1300 and 75000) culled from users of social media. Another paper using language, Mairesse and Walker 2006 claims that personality should be treated as an ordinal value, and as such uses rank error, which they minimize to between .26-.39 (where 0.5 is random) on a much smaller sample of participants (96).

The papers on device usage vary in their approach, Saati, Salem, and Brinkman 2005 uses Spearman's correlations (a correlation that considers rank explicitly), on 30 traits/subtraits with 16 participants and correlations of .46-.73. Khan et al. 2008 uses Pearson correlations, on 30 traits/subtraits with 26 participants and reported correlations between .4 and .6, and finally Oliveira et al. 2011 uses MSE, which can easily be translated to and interpreted as Pearson correlations at which point they range between .55 - .66 for 5 traits with 39 participants. The first two papers are more or less useless from their tiny sample size compared to the number of features they measured without proper corrections for multiple testing. The numbers from Oliveira et al. 2011 are slightly more interesting, and the correlations are extremely high. It is worth noting that to achieve this they use both mobile phone usage and social network analysis on the users mobile network data.

The clear divide is between treating this as a rank or continuous-value problem, which somewhat has to do with the accuracy of the test in placing individuals on a score that is representative and can be compared to different populations (in which case it is a continuous-value problem), versus seeing the test as an activity performed within a certain context that can only be compared within that context (in which case it could make more sense to consider this as a rank problem). In any case, an easy recommendation at this stage is to continue to consider both and assess the sensitivity of all prediction tasks to these two classes of error formulations.

2 How Good is Good

This subject is really only addressed in the two social-media language processing papers, Arnoux et al. 2017 was the foundation of IBM's Personality Insights, while Schwartz et al. 2013 was highly referenced as the "previous state of the art" by the former, and for that reason they went with the same error measurement and consider their performance "good."

Schwartz et al. 2013 references two texts from the psychology literature in order to justify their correlation scores as being "close to as good as possible". These two texts, Roberts et al. 2007 and Meyer et al. 2001 are both meta-analyses of dozens of psychology studies. They report that correlations between personality traits and behavior never exceed the 0.3-0.4 range. Another interesting fact that Meyer et al. 2001 reports is that correlations between self-reported behavioral problems (not traits specifically) have a similar relatively low level of correlation with reports of behavioral problems by experts or family members.

While Schwartz et al. 2013 introduces this lack of strong correlation between psychology traits and behavior as an upper-limit to how well prediction could be accomplished, this does not necessarily apply to our case, as we not looking at correlations with real-world outcomes but rather correlations with constructed tests. It should, however, be taken seriously into account as we consider different possible ways in which we gather data. If correlations for real-world behavior never seem to go beyond 0.40 at best, then we would hardly expect, from any single source of data of natural behavior, to get much better than that.

Clearly, it is important to consider the accuracy of our "ground truth.". The lack of correlation between self-reports and other forms of psychological tests in Meyer et al. 2001 should be investigated more closely in terms of personality traits specifically. Correctly measuring the variance or uncertainty in our measure of this ground truth is necessary in order to decide if a proxy is "good enough."

In many of these papers, the validity of the ground truth, the big five personality test that is administered, is not explored beyond reporting the Cronbach's alpha on their test results and claiming the alpha is "high" (in these cases, .8 - .95), and then going on to treat the score as a true value, rather than a random variable of any sort.

Cronbach's alpha, primarily used in this context of psychology tests, is simply a way to compare the variance of a trait to the variance of the individual survey questions that go into the trait. The idea is that if there is high variance in the questions that go into a trait, that trait is less accurately measured.

This idea can be extended in a number of ways in our circumstances, assuming the underlying logic is sound. For example, we can look at the within-trait variance per-student as a measure of uncertainty for a certain students score in a given trait. Such an idea naturally gives rise to a maximum-likelihood formulation where the variance of the noise expected for each prediction is based on the variance of a student's responses. Naturally this destroys the simplicity of measuring correlations for an objective, which assumes, like R^2, identical noise assumptions for every observation, but opens up ways for us to formally consider the uncertainty in our ground truth.

Even without such an approach, it is worth exploring ways in which we can use Cronbach's alpha as a measure of uncertainty in our ground truth, and how that leads to an answer to the question: "how good is good?"

References

[1] Pierre-Hadrien Arnoux et al. "25 Tweets to Know You: A New Model to Predict Personality with Social Media". In: (2017), pp. 25–28. arXiv: 1704.05513. URL: https://arxiv.org/ftp/arxiv/papers/1704/1704.05513.pdf%7B%5C%%7D5Cnhttp://arxiv.org/abs/1704.05513.

[2] Iftikhar Ahmed Khan et al. "Measuring personality from keyboard and mouse use". In: *Proceedings of the 15th European conference on Cognitive ergonomics the ergonomics of cool interaction - ECCE '08* July 2017 (2008), p. 1. DOI: 10.1145/1473018.1473066. URL: http://portal.acm.org/citation.cfm?doid=1473018.1473066.

[3] François Mairesse and Marilyn a. Walker. "Automatic recognition of personality in conversation". In: *Proceedings of the Human Language Technology Conference of the NAACL* June (2006), pp. 85–88. ISSN: 0749596X. DOI: 10.3115/1614049.1614071.

[4] Gj Meyer et al. "Psychological Testing and Psychological Assessment". In: *American Psychologist Copyright* 56.2 (2001), pp. 128–165. ISSN: 0003-066X. DOI: 10.1037//0003-066X.56.2.128. URL: http://psych.colorado.edu/%7B~%7Dwillcutt/pdfs/Meyer%7B%5C_%7D2001.pdf.

[5] Rodrigo de Oliveira et al. "Towards a psychographic user model from mobile phone usage". In: *CHI'11 Extended ...* (2011), pp. 2191–2196. DOI: 10.1145/1979742.1979920. URL: http://doi.acm.org/10.1145/1979742.1979920%7B%5C%%7D5Cnhttp://dl.acm.org/ft%7B%5C_%7Dgateway.cfm?id=1979920%7B%5C&%7Dtype=pdf%7B%5C%%7D5Cnhttp://dl.acm.org/citation.cfm?id=1979920.

[6] Brent W. Roberts et al. "The Power of Personality: The Comparative Validity of Personality Traits, Socioeconomic Status, and Cognitive Ability for Predicting Important Life Outcomes". In: *Perspectives on Psychological Science* 2.4 (2007), pp. 313–345. ISSN: 17456924. DOI: 10.1111/j.1745-6916.2007.00047.x. arXiv: 15334406.

[7] B. Saati, May Salem, and Willem-Paul Brinkman. "Towards customized user interface skins: investigating user personality and skin colour". In: *Hci 2005* 2 (2005), pp. 86–93.

[8] H. Andrew Schwartz et al. "Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach". In: *PLoS ONE* 8.9 (2013). ISSN: 19326203. DOI: 10.1371/journal.pone.0073791. arXiv: 1690219.1690245\x{fffd}\x{fffd}\x{fffd}.