CSE 250B Winter 2019 - Quiz 2

Feb 1, 2019 - 20 Minutes **Name:** **PID:**

---

(1) This is a closed book, closed notes exam. Switch off your cell phone and do not communicate with anyone other than an exam proctor.
(2) Start writing when instructed. Stop writing when your time is up.
(3) Remember that your work is graded on the quality of your writing and explanation as well as the validity of the mathematics.

Remember that in lecture, we talked about the statistical learning framework, where training, validation and test data are all independent samples drawn from the same underlying data distribution. State whether the statistical learning framework assumption applies in the following cases. In each case, if your answer is yes, explain why. If your answer is no, explain what is different between training and test data – μ (the marginal over x), η (the conditional distribution of $y|x$), or something else – and justify your answer.

(1) (5 points) Alice wants to build a classifier that detects if a patient's chest X-ray shows tuberculosis. She gathers a large training data set by collecting all patient chest X-rays from the UCSD hospital, and getting UCSD doctors to label them; this data is then used to build a classifier. The classifier is tested in the Mayo Clinic in New York.

The statistical learning framework does not hold – μ changes between training and test, but η does not. Whether a particular chest x-ray corresponds to tubercolosis or not still remains the same – whether the x-ray comes from a patient in San Diego or New York; however, the types and distribution of patients in the two locations may be different. For example, San Diego may have more patients with a certain type of tubercolosis.

(2) (5 Points) Bob wants to build a classifier that can predict whether a candidate would be a good engineer in his company. For this, he collects historical data of all applicants for engineers, along with who was hired; this is used to build a classifier, which is then tested on future job applications.

The statistical learning framework does not hold here either, as both μ and η are different between training and test. μ changes because the distribution of applicants and their properties changes over time. η is also different – whether a person is hired as an engineer or not often reflects biases of the hiring party, and is not always a good reflection of whether they turn out to be a good engineer or not.