WEEKLY PARTICIPATION 2: CONDITIONAL INDEPENDENCE

Often we can have a high number of potential features to be used in predicting our target $\boldsymbol{y}$, e.g. $\boldsymbol{x} \in \mathbb{R}^{1000}$, and a large number of these features may not be relevant to the prediction of the target.

(a) Let S be indices of some subset of the features, and $\boldsymbol{x}_S$ denote the corresponding random vector. Use the notion of independence to explain concisely with mathematical notation when the features $\boldsymbol{x}_S$ are irrelevant to predicting $\boldsymbol{y}$.

(b) More subtly, if we have a good subset of predictors $\boldsymbol{x}_G$ already, then we may say that a candidate set of features $\boldsymbol{x}_S$ doesn't add any additional value on top of $\boldsymbol{x}_G$ in predicting $\boldsymbol{y}$. Use the notion of conditional independence to explain concisely with mathematical notation when this happens.

Determining independence and conditional independence is useful for *feature selection*: the task of determining a reduced set of features for predicting the target. Feature selection helps reduce the computational effort in machine learning, and can lead to predictive models with lower variance. On a related note, conditional independence is crucial in building models that ignore spurious correlations, so generalize more reliably: recall the example in given in class where instead of learning a meaningful facial image recognition system, a model can mistakenly learn to recognize faces by the backgrounds they are often in front of in the training data set. That is an example of a spurious correlation; these can be filtered out using conditional independence checks.