Fairness and reliability in ML (COL865) **Minor II** **October 19, 2019**

Name: **Entry:** **Note:** Please answer only in the space provided.

---

1. Let us re-visit the three criteria of fairness, where Y is the outcome variable, R is the classifier score and Z is a sensitive attribute representing a protected category.

   (a) **Independence:** $R \perp Z$, or $P(R = 1|Z = a) = P(R = 1|Z = b)$.

   (b) **Separation:** $R \perp Z|Y$, or $P(R = 1|Y = 1, Z = a) = P(R = 1|Y = 1, Z = b)$ and $P(R = 1|Y = 0, Z = a) = P(R = 1|Y = 0, Z = b)$
   Note that $P(R = 1|Y = 1)$ and $P(R = 1|Y = 0)$ are the true and false positive rates, respectively.

   (c) **Sufficiency:** $Y \perp Z|R$, or $P(Y = 1|R = r, Z = a) = P(Y = 1|R = r, Z = b)$

   Comment on the following:

   (a) Sufficiency implies that the classifier is well calibrated, i.e., if the classifier identifies a set of people as having a probability r of constituting positive instances, then approximately a fraction r of this set should indeed be positive instances.

   (b) The strengths and limitations of each of these criteria of fairness, also taken together.

(c) Any two are in general impossible to achieve. Please provide the intuition, not the maths.

(d) All the three above are observational criteria. Suggest a criterion that is not observational.

2. Let us revisit the UC Berkeley admission problem from 1973. The overall acceptance

	Men		Women	
Department	Applied	Admitted (%)	Applied	Admitted (%)
A	825	62	108	**82**
B	520	60	25	**68**
C	325	**37**	593	34
D	417	33	375	**35**
E	191	**28**	393	24
F	373	6	341	**7**

rate for men was 44%, whereas for women it was only about 30%, suggesting the possibility of disparate impact. The department-wise data (see figure), however, showed a higher acceptance rate for women for most departments (please argue that this is not a mathematical surprise!). Apparently, Berkeley argued that

> "The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seems quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects."

(a) Argue that there is no way to conclude the debate from the given data.

(b) Suggest what could Berkeley have done at the time of admission to determine whether there was discrimination and prevent if possible. Assume that the admissions were based on a score.