KDG Background Reading

- What are Polytopes?

  a geometric object with flat sides.

- Kernel Density Estimation (KDE)

  - a non-parametric way to estimate the PDF of a random variable.
  - Let $(x_1, x_2, \ldots, x_n)$ be i.i.d. sampled drawn from a known distribution f at a given point x. We are interested in estimating the shape of the function f.

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where $K_h \rightarrow$ Kernel (a non-negative f^n)

$h \rightarrow$ Bandwith (a smoothing parameter)

- Performance Metrics

① Expected Caliberation Error (ECE)

Measures the expected difference between accuracy and confidence by grouping all samples (size N) into K bins & calculating

$$ECE = \sum_{i=1}^{K} \frac{|B_i|}{N} |acc_i - conf_i|$$

acc_i and $conf_i$ are accuracy & average confidence in the i-th bin & $|B_i|$ is the number of samples in bin B_i

* The pseudo-probabilities are class probabilities we get from the final layer of a NN.
  The pseudo-probability of the predicted class generally over-estimates the actual probability of getting a correct answer.

* If this over-estimation can be measured, it can be used to caliberate the NN such that it's pseudo-probabilities would match the actual probability of the classes.

  confidence

② **Cohen's Kappa**

$$K = \frac{P_o - P_e}{1 - P_e}$$

$P_o \rightarrow$ relative observed agreement among raters
$P_e \rightarrow$ hypothetical probability of chance agreement.

* Measures the agreement between 2 raters who each classify N times into C mutually exclusive categories.

* It is a quantitative measure of reliability for 2 raters that are rating the same thing, corrected for how often that the raters may agree by chance.

rater 2

		correct	incorrect
rater 1	correct	A	B
	incorrect	C	D

In A & D, the two raters are in agreement.

$$P_o = P(\text{agreement}) = \frac{A+D}{A+B+C+D} = \frac{\#\text{agreements}}{\text{Total}}$$

Expected probability that both would say correct ⇒

$$P(\text{correct}) = \frac{A+B}{A+B+C+D} \times \frac{A+C}{A+B+C+D}$$

Expected probability that both would say incorrect ⇒

$$P(\text{incorrect}) = \frac{C+D}{A+B+C+D} \times \frac{B+D}{A+B+C+D}$$

Pe = overall random agreement probability that they agreed on either yes or no.

$$P_e = P(\text{correct}) + P(\text{incorrect})$$

Then $$K = \frac{P_o - P_e}{1 - P_e}$$

Understanding KDN

* Understanding Polytopes formed by MLPs

- Partition & vote
- representation space learnt by DNNs is a partitioning of feature space in to a union of convex polytopes.

classical statistical formulation of classification problem ⇒

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \overset{iid}{\sim} F_{XY}$$

$T_n = \{(X_i, Y_i)\}_{i \in \{1, \ldots, n\}}$: training data

(X, Y) : to-be classified test observation & its true label

$X \in \mathbb{R}^d$, $Y \in \{0,1\}$

Learn a classification rule $g_n = g(\cdot\,; \tau_n)$ that maps feature vectors to class labels s.t. probability of misclassification $L(g_n) = P[g(X;\tau_n) \neq Y \mid \tau_n]$ is small.

Stone's Th$^{\underline{m}}$ for universally consistent classification:

a successful classifier can be constructed by

1) partitioning the input space into cells depending on n, such that the # training points in each cell goes to infinity but slowly in n

2) Estimating the posterior $\eta(x) = P[Y=1 \mid X=x]$ locally by voting based on the training class labels associated with the training feature vectors in cell $C(x) \subset \mathbb{R}^d$ in which the test observation falls.

Then $L(g_n) \to L^*$ almost surely for any F_{XY} where L^* is the Bayes optimal probability of misclassification.