Section 4A. Classification

Statistics for Data Science

Victor M. Preciado, PhD MIT EECS
Dept of Electrical & Systems Engineering
University of Pennsylvania
preciado@seas.upenn.edu

Classification Problem

- **Classification problem**:
  - Consider a discrete set $\mathcal{C}$, which contains K class labels. In this problem, the output variable Y is *qualitative* and takes values from $\mathcal{C}$
  - Our goal is to build a classifier $C(\mathbf{x})$ that assigns a class label in $\mathcal{C}$ to an input $\mathbf{x}$
  - *Examples*: Image classification, handwriting recognition, spam email detection, etc.
- **Generative model**:
  - In our analysis, we assume that points in our dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ are generated using the following two steps:
    1. Sample an input $\mathbf{x}_i \sim f_X$, where f_X is a marginal PDF
    2. Sample the corresponding output $y_i \in \mathcal{C}$ using this conditional PMF:

$$p_k(\mathbf{x}_i) = \Pr(Y = k | X = \mathbf{x}_i), \text{ for all } k \in \mathcal{C}$$

where $p_k(\mathbf{x})$ is called the *conditional class probability*...

Classification Problem: Numerical Example

- **Numerical example**: We generate $N = 100$ random samples according to the following distributions:

$$X \sim \text{Unif}\left(-2, 2\right) \text{ and } p_k\left(x\right) = \begin{cases} \frac{1}{1+e^{-2x}} & \text{for } k = 1 \\ \frac{e^{-2x}}{1+e^{-2x}} & \text{for } k = 0 \end{cases}$$

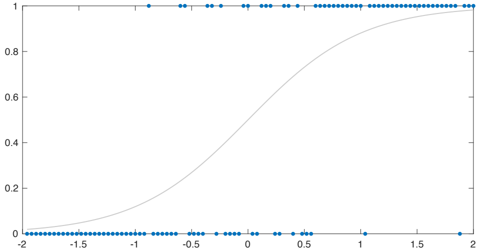

Figure: Scatter plot of our dataset (blue dots) and the function $1/\left(1+e^{-2x}\right)$ (gray line).

Penn Engineering