Exercise Sheet 4a

Exercise 1

Let x be a random variable. Consider the moment generating function

$$\phi_x : \mathbb{R} \to \mathbb{R}, \quad s \mapsto \mathbb{E}[\exp(sx)].$$

a) The exponential distribution with parameter $\lambda > 0$ is

$$p_\lambda(x) = \begin{cases} \lambda \exp(-\lambda x) & : \quad x \geq 0 \\ 0 & : \quad x < 0 \end{cases} .$$

Let $\lambda > s$. Show that

$$\phi_x(s) = \frac{\lambda}{\lambda - s}$$

for an exponentially distributed random variable x.

b) Show that $\phi_x(s)$ satisfies the following properties:

1. $\phi_x(0) = 1$
2. $\phi_x^{(k)}(0) = \mathbb{E}[x^k]$ for any integer $k \geq 0$

where $\phi_x^{(k)}(s)$ is the k^{th} derivative of ϕ_x at s.

c) Consider the independent random variables $x_1, \ldots, x_n$. Let $z = x_1 + \cdots + x_n$. Show that

$$\phi_z(s) = \prod_{i=1}^{n} \phi_{x_i}(s).$$

What is the form of $\phi_z(s)$ when the random variables x_i are also identically distributed?

d) Let $x_1, \ldots, x_n$ be i.i.d. Bernoulli random variables with parameter p. Let $z = x_1 + \cdots + x_n$ be the sum. Use $\phi_z(s)$ to show that $\mathbb{E}[z] = np$ and optionally $\mathbb{V}[z] = np(1-p)$.

Context: Why do we care about this exercise?

Exercise 2

In this exercise, you will apply Hoeffding's error bounds to a binary classification problem. For a fixed classifier f and a given $\delta > 0$, the error bounds are

$$R_n(f) - \sqrt{-\frac{\ln(\delta/2)}{2n}} \leq R(f) \leq R_n(f) + \sqrt{-\frac{\ln(\delta/2)}{2n}}$$

Notebook `exercise_2.ipynb` provides code to generate a ground-truth population $\mathcal{Z} \subseteq \mathcal{X} \times \mathcal{Y}$ of N training examples from a fixed distribution. Apply different classifiers to the following experiment:

1. Choose a fixed model f. For example, fit a model on a randomly chosen training set of size 100 from $\mathcal{Z}$.
2. Compute the true risk $R(f)$ over all examples from $\mathcal{Z}$.
3. Repeat for different test set sizes n:
   (a) Repeat T times:
      i. Sample a test set of size n from from $\mathcal{Z}$.
      ii. Compute the empirical risk $R_n(f)$ on the test set.
      iii. Compute Hoeffding's error bounds based on $R_n(f)$.
   (b) Average the error bounds over all T trials
4. Plot the true risk $R(f)$ and the average error bounds as a function of the test set size n.