Information, Entropy, and intro to Multivariate Normal

Submit a PDF of your answers to Canvas.

1. It's 1947 and you work at Bell Labs. You want to find a function $\mathcal{I}(\cdot)$ that represents the amount of information you learn from a random experiment. You come up with a list of requirements:

   - $\mathcal{I}(\cdot)$ should only be a function of the probability that outcome occurred: i.e, $\mathcal{I}(\cdot)$ is only a function of $\mathbb{P}(X = x)$, so that you can write $I(\mathbb{P}(X = x))$.

   - For independent X_1 and X_2, $\mathcal{I}(\mathbb{P}(X_1 = x_1, X_2 = x_2)) = \mathcal{I}(\mathbb{P}(X_1 = x_1)) + \mathcal{I}(\mathbb{P}(X_2 = x_2))$.

   - $\mathcal{I}(\cdot)$ is always non-negative.

   - $\mathcal{I}(\cdot)$ is decreasing in its argument.

   - $\mathcal{I}(1) = 0$.

   a) Argue why each of these items is a reasonable requirement for a notion of *information*.

   b) Show that $\mathcal{I}(\cdot) = \log_2\left(\frac{1}{\mathbb{P}(X=x)}\right)$ satisfies each of these requirements.

   c) **Optional.** Show that $\mathcal{I}(\cdot) = \log_k\left(\frac{1}{\mathbb{P}(X=x)}\right)$ is the only twice differentiable function that satisfies these properties.

2. Consider a classifier $\widehat{y} = f(\boldsymbol{x})$ where $\boldsymbol{x} \in \mathcal{X}^n$ is a random vector over a finite feature space $\mathcal{X}^n$. Show that the entropy of $\widehat{y} = f(\boldsymbol{x})$ is at most equal to the entropy of $\boldsymbol{x}$. *Hint:* start by expanding the joint entropy in two ways:

$$H(\boldsymbol{x}, \widehat{y}) = H(\widehat{y}) + H(\boldsymbol{x}|\widehat{y})$$
$$H(\boldsymbol{x}, \widehat{y}) = H(\boldsymbol{x}) + H(\widehat{y}|\boldsymbol{x})$$

   Bound the first equation, and simplify the second. Make sure to justify your steps.

3. Consider a random vector

$$\boldsymbol{x} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 3 \\ -1 \end{bmatrix}, \begin{bmatrix} 7 & 3 \\ 3 & 2 \end{bmatrix} \right)$$

a) Create a scatter plot of 1000 realizations of $\boldsymbol{x} \in \mathbb{R}^2$, with x_1 on the horizontal axis and x_2 on the vertical axis. Use `np.random.randn(2,1)` to create a single realization of a random vector $\boldsymbol{x}' \sim \mathcal{N}(0, \boldsymbol{I}) \in \mathbb{R}^2$, and then apply an appropriate linear transformation to so that $\boldsymbol{x}$ follows the specified distribution. You may also find `np.linalg.cholesky(B)` helpful.

b) Find an expression for the distribution of X_1 given that $X_2 = 0$.

c) Find the marginal distribution of X_2.

4. *Quadratic Discriminant Analysis.* Consider a random vector $\boldsymbol{x}$ that comes from one of two classes:

$$\boldsymbol{x}|Y = 0 \ \sim \ \mathcal{N}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 8 & 3 \\ 3 & 2 \end{bmatrix}\right)$$

$$\boldsymbol{x}|Y = 1 \ \sim \ \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}\right)$$

where each class occurs with probability $1/2$.

a) Design the MAP classification rule for Y given $\boldsymbol{x}$. Simplify your expression as much as possible.

b) Plot the decision boundary for your classifier along with a scatter plot that shows 1000 realization from each distribution above.

c) How many point are misclassified? What is the empirical risk of your classifier?