

## 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.

### SOLUTION:

- a) This question is more philosophical than most that we encounter in this class. Here are some arguments that can be made:
- The label that we assign to an outcome should not impact the amount of information that we learn from the outcome of a random experiment.
  - The amount of information we learn from two independent experiments should be the sum of the information from the individual experiments. Since the probability of two independent events is equal to their product, the information function should have the stated property.
  - This supports the notion that we are ‘learning’ information, which should be a positive quantity.
  - As the probability of an even becomes larger, it’s occurrence becomes less informative.
  - In the extreme, when an outcome is certain, we learn nothing.

- b) • Clearly this is true.
- $\mathbb{P}(X_1 = x_1, X_2 = x_2)) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)$  for independent  $X_1, X_2$ . Next,  $\log(ab) = \log(a) + \log(b)$ .
  - For  $a \in [0, 1]$ ,  $\log(1/a)$  is non-negative.
  - Since  $1/a$  is decreasing in its argument for  $a \in [0, 1]$ ,  $\log(1/a)$  is also decreasing. This can be verified by plotting or differentiating the function.
  - Note that  $\log(1) = 0$ .

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

$$\begin{aligned} H(\mathbf{x}, \hat{y}) &= H(\hat{y}) + H(\mathbf{x}|\hat{y}) \\ H(\mathbf{x}, \hat{y}) &= H(\mathbf{x}) + H(\hat{y}|\mathbf{x}) \end{aligned}$$

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

**SOLUTION:** We can bound the first equation by dropping the second term since it is always positive, and have  $H(\mathbf{x}, \hat{y}) \geq H(\hat{y})$ . In the second term,  $H(\hat{y}|\mathbf{x}) = 0$  since  $\hat{y}$  is a function of  $\mathbf{x}$ , i.e.,  $\hat{y} = f(\mathbf{x})$ . This gives the result:  $H(\hat{y}) \leq H(\mathbf{x})$ .

3. Consider a random vector

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

- a) Create a scatter plot of 1000 realizations of  $\mathbf{x} \in \mathbb{R}^2$ , 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  $\mathbf{x}' \sim \mathcal{N}(0, \mathbf{I}) \in \mathbb{R}^2$ , and then apply an appropriate linear transformation to so that  $\mathbf{x}$  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  $X_2$ .

**SOLUTION:**

- a) If  $\mathbf{x}' \sim \mathcal{N}(0, \mathbf{I}) \in \mathbb{R}^2$ , then  $\mathbf{A}\mathbf{x}' + \mu \sim \mathcal{N}(\mu, \mathbf{A}\mathbf{A}^T)$ . Hence, we need a matrix  $\mathbf{A}$  such that  $\mathbf{A}\mathbf{A}^T = \Sigma$ . Taking the Cholesky factorization of  $\Sigma$  gives an appropriate  $\mathbf{A}$ ; hence, `np.linalg.cholesky(Sigma)` returns such a matrix. To create random variables according to the specified normal distribution, we take  $\mathbf{A}\mathbf{x}' + \mu$ . See the associated notebook for code.

- b) We are interested in an expression for the conditional distribution of  $X_1$  given  $X_2 = 0$ . Adjusting the expression from lecture, we have

$$x_1|x_2 \sim \mathcal{N}(\Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) + \mu_1, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

hence, the conditional is

$$x_1|x_2 \sim \mathcal{N}\left(\frac{3}{2}(x_2 + 1) + 3, 7 - \frac{9}{2}\right)$$

When  $x_2 = 0$ , we have  $x_1|x_2 = 0 \sim \mathcal{N}\left(\frac{9}{2}, \frac{5}{2}\right)$ .

- c) By inspection,  $X_2 \sim \mathcal{N}(-1, 2)$ .

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

$$\begin{aligned} \mathbf{x}|Y=0 &\sim \mathcal{N}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 8 & 3 \\ 3 & 2 \end{bmatrix}\right) \\ \mathbf{x}|Y=1 &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0.1 \\ 0.1 & 1 \end{bmatrix}\right) \end{aligned}$$

where each class occurs with probability 1/2.

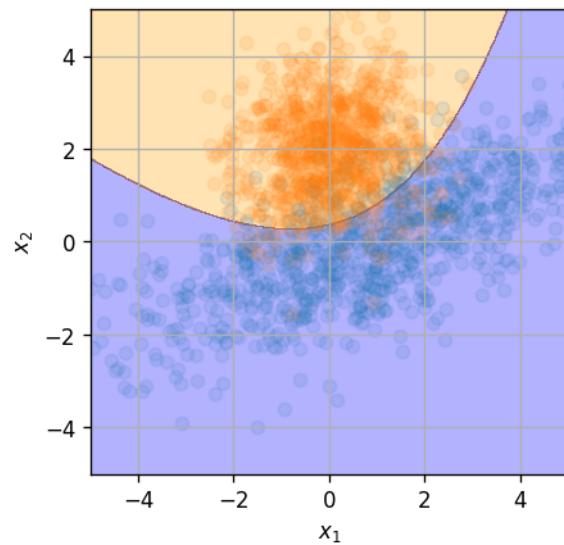
- a) Design the MAP classification rule for  $Y$  given  $\mathbf{x}$ . Simplify your expression as much as possible.
- b) Plot the decision boundary for your classifier along with a scatter plot that shows 1000 realization from each distribution above.
- c) How many point are misclassified? What is the empirical risk of your classifier?

### SOLUTION:

- a) We know the MAP decision boundary under a uniform prior has the form:  $\mathbf{x}^T \mathbf{B} \mathbf{x} + \mathbf{w}^T \mathbf{x} > c$  where

$$\begin{aligned} c &= \log(|\Sigma_0|) - \log(|\Sigma_1|) + \boldsymbol{\mu}_0^T \Sigma_0^{-1} \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^T \Sigma_1^{-1} \boldsymbol{\mu}_1 \\ \mathbf{w} &= 2(\Sigma_0^{-1} \boldsymbol{\mu}_0 - \Sigma_1^{-1} \boldsymbol{\mu}_1) \\ \mathbf{B} &= \Sigma_1^{-1} - \Sigma_0^{-1}. \end{aligned}$$

b) See plot below and associated notebook.



c) Approximately 8% of the points were misclassified.