CAS ETH Machine Learning in Finance and Insurance. Mini-exercises - Lecture 1.

Andrea Ferrario

Version of March 6, 2024

1 Probability spaces and random variables

1. Suppose you flip a fair coin three times. What is the probability of getting exactly two heads?

2. Let us toss a coin twice. What is the sample space? What is the event "the first toss is heads"?

3. Consider a game where you roll a fair six-sided die. If you roll an even number, you win a prize. List the sample space of this experiment and calculate the probability of winning a prize.

4. Suppose you flip two fair coins simultaneously. List the sample space for this experiment and calculate the probability of getting at least one head.

5. A password is formed by randomly selecting three letters from the alphabet (A-Z) with replacement. What is the sample space for this experiment? Calculate the probability that all three letters are the same.

6. Let X be a discrete random variable representing the sum of two fair six-sided dice rolls. Compute the expected value $\mathbb{E}[X]$.

7. Suppose Y is a continuous random variable with a uniform distribution over the interval $[0, 2]$. Find $P(0.5 \leq Y \leq 1.5)$.

8. Let Z be a random variable that follows a normal distribution with a mean of 10 and a standard deviation of 2. Find $P(Z > 12)$.

9. A random variable X follows a Poisson distribution with rate parameter $\lambda = 3$. Find the probability that X is less than 2.

2 Statistical learning

1. What is statistical learning, and why is it important in the context of data analysis and predictive modeling?

2. Describe the difference between supervised and unsupervised learning. Provide an example of each.

3. Explain the concept of overfitting in statistical learning and how it can affect the performance of a model.

4. Compare and contrast classification and regression in the context of statistical learning. Provide an example where each would be appropriately applied.

5. What is meant by the "true model" in statistical learning, and why is it considered unobservable? What would it happen to machine learning modeling if we knew the true model?

6. Discuss the role of error in statistical learning.

7. What role does the loss function play in the construction of statistical learning models? Provide examples of common loss functions used in regression and classification.

8. What does it mean to "learn" or "train" a machine learning model?

9. Why is checking the performance of a machine learning model on training data not enough to justify its use or deployment?

10. Explain the concept of "model complexity" heuristically. How does it relate to the performance of a machine learning model?

3 Linear regression

1. Explain what linear regression is in terms of statistical learning.

2. Given the training dataset $\{(x_i, y_i)\}_{i=1}^n$ and a linear regression model in d variables, write down the least squares loss function.

3. Given the training dataset $\{(x_i, y_i)\}_{i=1}^n$ and a linear regression model in d variables, write down the least squares loss function with LASSO, ridge and elastic net regularization terms.

4. Derive the normal equations for the linear regression. Explain how they are used to estimate the coefficients of a linear regression model.

5. Derive the formula for the intercept and the slope of a linear regression model in one variable analytically.

6. Assume a linear regression model with two predictors X_1 and X_2, and response variable Y. Given the matrices:

$$\mathbf{A} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

Write the normal equations explicitly and describe how you would solve them to estimate the three coefficients of the model. How do the normal equations look like with ridge regularization?

7. (Extra) In Python, create a synthetic dataset for a linear regression model with one predictor variable. Implement a function that computes the regression coefficients analytically, given your synthetic data. Verify your results by computing the coefficients using the function `LinearRegression()` in `sklearn`.