PC3: probabilities, bias-variance trade-off and decision trees

APM_51434_EP
École Polytechnique

In this PC, we will work on exercises covering probabilities, conditional probabilities, expectations, conditional expectations, and variances. We will then apply these concepts to the bias-variance decomposition, learning how to break down the error of a model into its different components, together with the link with the size of the hypothesis class. Finally, we will study the "expressivity" of decision trees and discuss how their bias and variance are impacted by the tree's depth.

We first give a few reminders about discrete random variables. Let X, Y be two discrete random variables taking values in a countable set $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$, we recall :

— *probability mass function (pmf)* : $P(X = x) = p_X(x)$, with $\sum_{x\in\mathcal{X}} p_X(x) = 1$.

— *joint and conditional probability pmfs* : $p_X(x) = \sum_{y\in\mathcal{Y}} p_{X,Y}(x, y)$, $p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$.

— *law of total probability* : $p_Y(y) = \sum_{x\in\mathcal{X}} p_{Y|X}(y|x)p_X(x)$.

— *(conditional) expectation* : $E[X] = \sum_{x\in\mathcal{X}} x\, p_X(x)$, $E[g(Y) \mid X = x] = \sum_{y\in\mathcal{Y}} g(y)\, p_{Y|X}(y|x)$.

Conditional Probability and Expectation

Exercice 3.1 (The "medical test paradox")**.** A disease affects 1% of the population. We define D and T two random variables, with $D = 1$ if a person has the disease and $D = 0$ if they do not ; moreover $T = 1$ denotes a positive test result and $T = 0$ a negative test result.

We study a test which has a sensitivity $P(T = 1 \mid D = 1) = 0.99$ (probability of true positive), and a specificity $P(T = 0 \mid D = 0) = 0.95$ (probability of true negative).

1. Starting from the definition of conditional probability, derive Bayes' theorem :
$$P(D = 1 \mid T = 1) = \frac{P(T = 1 \mid D = 1)P(D = 1)}{P(T = 1)}.$$
2. Compute $P(T = 1)$ using the law of total probability and evaluate $P(D = 1 \mid T = 1)$.
3. Explain mathematically why the posterior probability is small when $P(D = 1)$ is low.
4. Show that
$$\frac{P(D = 1 \mid T = 1)}{P(D = 0 \mid T = 1)} = \frac{P(T = 1 \mid D = 1)}{P(T = 1 \mid D = 0)} \times \frac{P(D = 1)}{P(D = 0)}.$$
Interpret this as an update of prior odds by the likelihood ratio.

Exercice 3.2 (Conditional expectation)**.** Let X, Y be discrete random variables taking values in finite sets $\mathcal{X}, \mathcal{Y}$ with $\mathcal{X} \subseteq \mathbb{R}$. The conditional expectation $E[X \mid Y]$ is
$$E[X \mid Y = y] = \sum_{x\in\mathcal{X}} x\, P(X = x \mid Y = y).$$

1. Prove the tower property $E[E[X \mid Y]] = E[X]$.
2. Let $g\colon \mathcal{Y} \to \mathbb{R}$, show that $E[Xg(Y)] = E[E[X \mid Y]g(Y)]$.

Exercice 3.3 (Bonus : Conditional variance decomposition)**.** Let X, Y be square-integrable discrete random variables taking values in finite sets $\mathcal{X}, \mathcal{Y}$. Define
$$\mathrm{Var}(X \mid Y) = E[(X - E[X \mid Y])^2 \mid Y] = \sum_{x\in\mathcal{X}} (x - E[X \mid Y])^2 P(X = x \mid Y).$$

1. Write the conditional variance as a finite sum over the values that X can take.
2. Prove the law of total variance :

$$\mathrm{Var}(X) = E[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(E[X \mid Y]).$$

3. Let $X = Y + \varepsilon$, with Y and ε taking values in $\{0, 1\}$ independent. What does the law of total variance give ?

Bias-Variance Decomposition

Exercice 3.4. Let Z be a square-integrable real-valued random variable. Let c be a constant. Consider the quantity

$$E\big[(Z - c)^2\big].$$

1. Choose $c = E[Z]$. What is $E[(Z - c)^2]$ in this case ?
2. If $c \neq E[Z]$, how does $E[(Z - c)^2]$ compare to the previous case ? How can we interpret the additional term ?
3. Suppose now that C is a random variable (possibly dependent on X), what can you say about $E[(Z - C)^2]$? Use conditional expectations and variances.

Exercice 3.5 (Bias-Variance Decomposition)**.** Let (X, Y) be a random pair, where $X \in \mathcal{X}$ are features and $Y \in \mathbb{R}$ is a continuous target. Let $\hat{f}_D(x)$ be a random estimator of Y at input x, depending on a random dataset D that is independent of (X, Y), and let $f^*(x) = E[Y \mid X = x]$ denote the true conditional mean. Define the conditional mean-squared error

$$\mathcal{E}(x) = E\big[(Y - \hat{f}_D(x))^2 \mid X = x\big].$$

1. Let $\sigma^2(x) = \mathrm{Var}(Y \mid X = x)$. Show that $\mathcal{E}(x)$ can be decomposed as

$$\mathcal{E}(x) = \sigma^2(x) + (E[\hat{f}_D(x)] - f^*(x))^2 + \mathrm{Var}[\hat{f}_D(x)],$$

   Hint : add/subtract $f^*(x)$ in the square in $\mathcal{E}(x)$ and expand.
2. Identify which terms of this decomposition are *bias* and *variance* terms.
3. Discuss how the bias and variance terms typically behave for (i) a small hypothesis class, (ii) a large hypothesis class ?

Decision Trees and VC Dimension

Exercice 3.6 (From "Understanding Machine Learning", Section 18.6, Exercise 1)**.**

1. Show that any binary classifier $h : \{0, 1\}^d \to \{0, 1\}$ can be implemented as a decision tree of height at most $d + 1$, with internal nodes testing coordinates of the form "$(x_i = 0)$?" for some $i \in \{1, \ldots, d\}$.
2. Conclude that the VC dimension of the class of decision trees over the domain $\{0, 1\}^d$ is 2^d. Explain the connection with overfitting.
3. What is the link between the tree depth and the bias–variance decomposition ? Discuss the case of trees of depth 1 (decision stumps) and of depth d.

Exercice 3.7 (Bonus : Sub-optimality of ID3)**.** Let $\{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\{0, 1\}^d, \mathbb{R})^n$ be a dataset with binary features. Does ID3 with pre-determined maximal depth always find a tree that distinguishes all the points ?

Hint : consider $\{((1, 1, 1), 1), ((1, 0, 0), 1), ((1, 1, 0), 0), ((0, 0, 1), 0)\}$ *with maximum depth* 2.