Johannes Gutenberg-Universität Mainz
Institut für Informatik
Data Mining
Prof. Dr. Stefan Kramer

Machine Learning - Sheet 6

01.06.2017
Deadline: 08.06.2017 - 23:55

Task 1: Random Variables and Bayes Rule *(6 Points)*

The goal of this task is to better understand the notation used during the derivation of the Naive Bayes algorithm.

First, some definitions. Let Ω be a finite set, and let $\mathbb{P}[\{\bullet\}]$ be a probability mass function on Ω. If C is some other set, and $Z : \Omega \to C$ some function, then we call Z a *C-valued random variable.* For any subset $A \subseteq C$, we define the event

$$\{Z \in A\} := \{\omega \in \Omega | Z(\omega) \in A\}.$$

As a special case, if $c \in C$ is a single element, we write:

$$\{Z = c\} := \{Z \in \{c\}\} \equiv \{\omega \in \Omega | Z(\omega) = c\}.$$

When we apply $\mathbb{P}[\bullet]$ to such events, we usually suppress the curly braces:

$$\mathbb{P}[Z \in A] := \mathbb{P}[\{Z \in A\}] \qquad \mathbb{P}[Z = c] := \mathbb{P}[\{Z = c\}].$$

Moreover, we often replace the set operations $\cup$ and $\cap$ by the corresponding logical operations $\vee$ and $\wedge$ (often, the $\wedge$ is replaced by just a single comma).

As an example, consider a spam/ham-classification task on a tiny language, in a tiny universe, where all emails consist of exactly two characters. Let $V := \{a, b\}$ be the vocabulary. Let $K := \{\ominus, \oplus\}$ be class labels for spam and "ham", respectively. Suppose that in the entire history of our tiny model universe, only eight emails have been sent (and then either read, or marked as spam):

$$aa\oplus, aa\oplus, ab\oplus, ab\oplus, ab\ominus, ba\oplus, bb\ominus, bb\ominus$$

(we have suppressed parentheses and commas in the tuples). The information from the above list can be summarized as a finite set with a probability mass function on it:

$$\Omega := \{aa\oplus, ab\oplus, ab\ominus, ba\oplus, bb\ominus\} \subsetneq L^2 \times K, \qquad \mathbb{P}[\{\omega\}] := \begin{cases} 0.25 & \text{for} \quad \omega = aa\oplus \\ 0.25 & \text{for} \quad \omega = ab\oplus \\ 0.125 & \text{for} \quad \omega = ab\ominus \\ 0.125 & \text{for} \quad \omega = ba\oplus \\ 0.25 & \text{for} \quad \omega = bb\ominus \end{cases}.$$

Define random variables

$$\begin{array}{ll} W_1 : \Omega \to L & W_1(xyz) := x \\ W_2 : \Omega \to L & W_2(xyz) := y \\ Class : \Omega \to K & Class(xyz) := z. \end{array}$$

With the above notation, we can, for example, write: $\mathbb{P}[W_2 = b] \equiv \mathbb{P}[\{ab\oplus, ab\ominus, bb\ominus\}] = 0.625$.

(1) Explicitly write out all elements of the following events. Briefly explain how each event can be interpreted.

- $\{W_1 = a\}$
- $\{W_2 = a\}$
- $\{W_1 = b\} \cap \{W_2 = a\}$
- $\{W_2 \in \{a, b\}\}$
- $\{Class = \ominus\}$
- $\{W_2 = b\} \cap \{Class = \ominus\}$

Example: $\{W_2 = b\} = \{ab\oplus, ab\ominus, bb\ominus\}$ = "Second character is 'b' ".

(2) Compute the following probabilities. Explain in words what each probability means.

- $\mathbb{P}[W_1 = a \wedge W_2 = b]$
- $\mathbb{P}[Class = \ominus]$
- $\mathbb{P}[Class = \oplus]$
- $\mathbb{P}[W_1 = a \wedge W_2 = b \,|Class = \ominus]$
- $\mathbb{P}[W_1 = a \wedge W_2 = b \,|Class = \oplus]$
- $\mathbb{P}[Class = \ominus \,|W_1 = a \wedge W_2 = b]$

(3) Explain why it is immediately obvious that

$$\mathbb{P}[Class = \ominus \,|W_1 = a \wedge W_2 = b] = \frac{\mathbb{P}[W_1{=}a \wedge W_2{=}b|Class{=}\ominus]\mathbb{P}[Class{=}\ominus]}{\mathbb{P}[W_1{=}a \wedge W_2{=}b|Class{=}\ominus]\mathbb{P}[Class{=}\ominus]+\mathbb{P}[W_1{=}a \wedge W_2{=}b|Class{=}\oplus]\mathbb{P}[Class{=}\oplus]}$$

must hold (without actually calculating anything). Now plug in the numbers from the previous subtask, verify that the equality really does hold.

(4) Find three events $D, h^+, h^- \subseteq \Omega$ such that the above equation syntactically unifies with

$$\mathbb{P}\left[h^- \,|D\right] = \frac{\mathbb{P}[D\,|h^-]\,\mathbb{P}[h^-]}{\mathbb{P}[D\,|h^-]\,\mathbb{P}[h^-] + \mathbb{P}[D\,|h^+]\,\mathbb{P}[h^+]}$$

(5) Now set $h := h^+$, and verify that

$$\mathbb{P}[h\,|D] = \frac{\mathbb{P}[D\,|h]\,\mathbb{P}[h]}{\mathbb{P}[D]}$$

also holds. What is the meaning of this equation in the context of spam detection?

Task 2: Mean Squared Error *(7 Points)*

The goal of this exercise is to understand the relationship between the mean squared error, the variance, and the bias.

For a real-valued random variable Z (as defined in Exercise 1), the *expected value* and the *variance* are defined as:

$$\mathbb{E}\left[Z\right] := \sum_{w \in \Omega} Z(w)\mathbb{P}\left[\{\omega\}\right] \qquad Var\left[Z\right] := \mathbb{E}\left[Z^2\right] - \mathbb{E}\left[Z\right]^2 .$$

Alternatively, with $Im(Z) := \{x \in \mathbb{R} \mid \exists\, \omega \in \Omega \text{ s.th. } Z(\omega) = x\}$ and the notation from Exercise 1, the expected value can be rewritten as:

$$\mathbb{E}\left[Z\right] := \sum_{x \in Im(Z)} x \cdot \mathbb{P}\left[Z = x\right] .$$

Observe that every $c \in \mathbb{R}$ can be interpreted as a constant function from Ω to $\mathbb{R}$, and therefore as a "random variable" (even though it's not random at all). With some abuse of notation, it holds:

$$\mathbb{E}\left[c\right] = c.$$

Show the following statements:

(1) If Z_1, Z_2 are two $\mathbb{R}$-valued random variables, and $\alpha, \beta \in \mathbb{R}$ some scalar factors, then it holds:

$$\mathbb{E}\left[\alpha Z_1 + \beta Z_2\right] = \alpha\mathbb{E}\left[Z_1\right] + \beta\mathbb{E}\left[Z_2\right] .$$

(This is the *linearity*, it allows you to rewrite $\mathbb{E}\left[\alpha Z\right] = \alpha\mathbb{E}\left[Z\right]$, $\mathbb{E}\left[A + B\right] = \mathbb{E}\left[A\right] + \mathbb{E}\left[B\right]$ etc.).

(2) Let $c \in \mathbb{R}$ be some constant, and Z a real-valued random variable. It holds:

$$\mathbb{E}\left[(Z - c)^2\right] = Var\left[Z\right] + \mathbb{E}\left[Z - c\right]^2 .$$

Hint: use $(a - b)^2 = a^2 - 2ab + b^2$, apply linearity, "add zero".

(3) Now assume that there are some finite sets X, Y with $Y \subset \mathbb{R}$, and that the true relationship between the instances $x \in X$ and the target values $y \in Y$ is given by a function $f : X \to Y$, so that $y = f(x)$ holds for all valid (x, y). A user samples a training set T of a fixed size n and uses a regression algorithm r to generate the regression function $r_T : X \to Y$ that is supposed to resemble the true relationship f. She then applies the generated function to a fixed test instance x_0 and obtains $y_0 := r_T(x_0)$. Since the sampled data T can be noisy, we decide to model $r_T(x_0)$ as a real-valued random variable. The *expected mean squared error at the point* x_0 of this prediction is given by

$$\text{MSE}(f, r_T, x_0) := \mathbb{E}\left[\left(r_T(x_0) - f(x_0)\right)^2\right] .$$

We define the *bias* as

$$\text{Bias}(f, r_T, x_0) := \mathbb{E}\left[r_T(x_0) - f(x_0)\right] .$$

Show that it holds:

$$\text{MSE}(f, r_T, x_0) = Var\left[r_T(x_0)\right] + \left(\text{Bias}(f, r_T, x_0)\right)^2 .$$

Hint: use (2).

Task 3: Logistic Regression with Nonlinear Features *(7 Points)*

The goal of this exercise is to see how logistic regression can be used to classify points in the plane, even in cases when the points are not linearly separable.

Run the attached script `ml17-exercise-06-linearRegression.py`. You will see three point clouds, which belong to two classes. These two classes are obviously not linearly separable.

(1) Implement the `gaussianRbf` function. This function takes coordinates of a central point $c = (c_x, c_y)$ (`centerX` and `centerY` in code), a squared sigma σ^2 (`sigmaSquared`), and coordinates of a point $p = (p_x, p_y)$. It is supposed to return

$$\phi_{c,\sigma^2}(p) := \exp\left(-\frac{\|p-c\|^2}{2\sigma^2}\right).$$

(2) Modify the function `extractFeatures`. Currently, it takes coordinates of a point (x, y), and then simply returns the coordinates, without modifying them. You should replace `[x, y]` by an array of non-linear features, computed from x and y.

- Try out various multivariate polynomials, for example `[x * x, x * y]`.
- Try out the `gaussianRbf` as nonlinear feature. Experiment with various settings for center and sigma.

Try to come up with at least three different ways to separate the points.

(3) Briefly discuss your results.