Machine Learning for Applied Physics and High Energy Physics

Lecture 2 Probability: univariate and multivariate models

2.1 Recap on Lecture 1

Machine Learning has three main ingredients

- EXPERIENCE or the data set. In the case of supervised learning, this is a collection of pairs features-labels $\mathcal{D} = \left\{ (\vec{x}_n, \vec{y}_n) \right\}_{n=1}^{N}$ with N the sample size and $D = \dim(\vec{x})$ the dimensionality of the data set.
- TASK or the mapping f (in supervised machine learning) such that $f: \mathcal{X} \longrightarrow \mathcal{Y}$. The task depends on the problem (CLASSIFICATION, REGRESSION).
- PERFORMANCE or the loss function $l(\vec{y}, f(\vec{x}, \vec{\vartheta}))$. Sometimes this is also called cost function.

Concerning the task, this course will focus on

SUPERVISED LEARNING UNSUPERVISED LEARNING

CLASSIFICATION REGRESSION CLUSTERING DIMENSIONALITY REDUCTION

2.2. Definitions of probability

This lecture is about generalities on probability. "Probability theory is nothing but common sense reduced to calculation". [Pierre Laplace, 1812].

1 FREQUENTIST PROBABILITY. The probability of an event is the limit of its relative frequency in many trials. Example: if we toss a coin many times we expect it to lead heads about half of the times.

2 BAYESIAN PROBABILITY. Probability is interpreted as a reasonable expectation representing a state of knowledge or as quantification of personal belief.

In this course, we adopt the second definition. Bayesian methods are characterised by concepts and procedures as follows

- The use of random variables to model ALL sources of uncertainty in statistical models:
  - aleatoric uncertainty (from data)
  - epistemic uncertainty (from model)
- The need to determine the prior probability distribution taking into account the available (prior) information.
- The sequential use of Bayes' theorem: as more data become available, calculate the posterior distribution using Bayes' theorem; subsequently the posterior distribution becomes the next prior.
- While for the frequentist a hypothesis is a proposition (which must be either TRUE or FALSE) in Bayesian statistics the probability that can be assigned to a hypothesis can also be in a range from 0 to 1 if the true value is uncertain

2.3 Bayes' theorem

Given an unknown quantity H and a set of known data $Y = y$ (H and Y both denote random variables) the probability of observing $H = h$ given the data is:

$$\underbrace{p(H=h / Y=y)}_{\text{POSTERIOR}} = \frac{\overbrace{p(H=h)}^{\text{PRIOR}} \, \overbrace{p(Y=y / H=h)}^{\text{LIKELIHOOD}}}{\underbrace{p(Y=y)}_{\text{MARGINAL LIKELIHOOD}}}$$

The theorem follows from the identity

$$p(h/y)\,p(y) = p(h)\,p(y/h)$$

which is the definition of conditional probability. Note that

$$p(Y=y) = \sum_{h'\in\mathcal{H}} p(H=h')\,p(Y=y/H=h') = \sum_{h'\in\mathcal{H}} p(H=h', Y=y)$$

Example: Let us suppose to have been infected by COVID-19. You take an antigenic test and you want to use its result to determine if you are infected or not. The SPECIFICITY of the test, i.e. the probability of being non-infectious if the test is negative (aka true negative) is 97,5%. The SENSITIVITY of the test, i.e. the probability of being infectious if the test is positive (aka true positive) is 87,5%. The prevalence of the disease is 10%. What's the probability of being actually infectious if the test is positive? And if it's negative? How do results change if the prevalence is 1%?

We call H the random variable such that

$$\begin{cases} h=1 & \text{infectious} \\ h=0 & \text{non-infectious} \end{cases}$$

We call Y the random variable such that

$$\begin{cases} y=1 & \text{positive test} \\ y=0 & \text{negative test} \end{cases}$$

Therefore, I want to calculate

$$p(H=1/Y=1) = \frac{p(H=1)\,p(Y=1/H=1)}{p(Y=1)} = 0,795 = 79,5\%$$

$$= \frac{p(H=1)\,p(Y=1/H=1)}{p(Y=1/H=1)\,p(H=1) + p(Y=1/H=0)\,p(H=0)} = \frac{0,1\cdot 0,875}{0,875\cdot 0,1 + (1-0,975)(1-0,1)}$$

$$p(H=1/Y=0) = \frac{p(H=1)\,p(Y=0/H=1)}{p(Y=0/H=1)\,p(H=1) + p(Y=0/H=0)\,p(H=0)}$$

$$= \frac{0,1\cdot(1-0,875)}{(1-0,875)\cdot 0,1+0,975\cdot(1-0,1)} = 0,014 = 1,4\%$$

If the prevalence is 1%

$p(H=1/Y=1) = 26\%$ $\qquad$ $p(H=1/Y=0) = 0,13\%$

2.4 Some definitions and properties

Given two events A and B (which correspond to random variables), one defines

$\Pr(A \cap B) = \Pr(A)\cdot\Pr(B) = \Pr(A,B)$ JOINT PROBABILITY

$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$ UNION PROBABILITY

$\Pr(B/A) = \dfrac{\Pr(A,B)}{\Pr(A)}$ CONDITIONAL PROBABILITY

$\Pr(A, B/C) = \Pr(A/C)\cdot\Pr(B/C)$ CONDITIONAL IDEPENDENCE

Given a random variable X, one defines

$p(x) \stackrel{\Delta}{=} \Pr(X = x)$ PROBABILITY MASS FUNCTION (PMF) (for a discrete random variable)

$P(x) \stackrel{\Delta}{=} \Pr(X \leq x)$ CUMULATIVE DISTRIBUTION FUNCTION (CDF (for a continuous random variable)

$p(x) \stackrel{\Delta}{=} \dfrac{d}{dx} P(x)$ PROBABILITY DENSITY FUNCTION

It follows that

$$\Pr(a < X < b) = \int_a^b dx\, p(x) = P(b) - P(a)$$

If the CDF P is strictly monotonically increasing, then it has an inverse P^{-1}, called quantile. In particular $P^{-1}(q)$ is the value x_q such that $\Pr(X \leq x_q) = q$. This is called the q'th quantile of P.

$$p(Y=y/X=x) = \frac{p(X=x, Y=y)}{p(X=x)} \quad \text{or} \quad p(x,y) = p(x)\,p(y/x)$$

This is the CONDITIONAL DISTRIBUTION. The chain rule is

$$p(\vec{x}_1, D) = p(\vec{x}_1)\,p(x_2/\vec{x}_1)\,p(x_3/\vec{x}_1, x_2)\ldots p(x_D/\vec{x}_1, x_{D-1})$$

Given two random variables X and Y and a third random var. Z

$$p(X,Y) = p(X)\,p(Y) \qquad \text{MARGINAL INDEPENDENCE}$$

$$p(X,Y,Z) = p(X/Z)\,p(Y/Z) \quad \text{CONDITIONAL INDEPENDENCE}$$

2.5 Moments of a distribution

- MEAN $\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x\,p(x)\,dx \quad \Big(= \underbrace{\sum_{x\in\mathcal{X}} x\,p(x)}_{\text{for a discrete variable}}\Big) = \mu$

The mean is linear in X: $\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$

- VARIANCE $\mathbb{V}[X] \triangleq \mathbb{E}\left[(X-\mu)^2\right] = \int_{\mathcal{X}} (x-\mu)^2 p(x)\,dx$

$$= \int_{\mathcal{X}} x^2 p(x)\,dx + \mu^2 \underbrace{\int_{\mathcal{X}} p(x)\,dx}_{1} - 2\mu \underbrace{\int_{\mathcal{X}} x\,p(x)\,dx}_{\mu} = \mathbb{E}[X^2] - \mu^2$$

Properties

$$\mathbb{V}[aX+b] = a^2\,\mathbb{V}[X] \qquad \mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}[X_i] \text{ for ind. variables}$$

• MODE

$$x^* = \arg\max_x p(x)$$ (most likely value; may not be unique)

The moments do not capture all the information about the PDF. Example: the Datasaurus Dozen.

2.6. The Gaussian distribution

CDF : $$P(x) \equiv \Phi(y;\mu,\sigma^2) \triangleq \int_{-\infty}^{y} \mathcal{N}(z/\mu,\sigma^2)\,dz = \frac{1}{2}\left[1+\operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right]$$

where $z = \frac{y-\mu}{\sigma}$ and $\operatorname{erf}(u) \triangleq \frac{2}{\sqrt{\pi}}\int_0^u e^{-t^2}dt$

PDF : $$p(x) \equiv \mathcal{N}(z/\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

$$\mathbb{E}[Y] = \int_y y\,p(y)\,dy = \mu \qquad \mathbb{V}[Y] = \int_y (y-\mu)^2 p(y)\,dy = \sigma^2$$

The Gaussian distribution is the most widely used in ML

- It has only two parameters which are easy to interpret
- The central limit theorem tells us that sums of independent random variables have an approximately Gaussian distribution
- The Gaussian distribution makes the least number of assumptions
- If two random variables are described by a Gaussian distribution, then the random variable, obtained as their sum, is described by a Gaussian distribution.

$$x_1 \sim \mathcal{N}(\mu_1,\sigma_1^2) \qquad x_2 \sim \mathcal{N}(\mu_2,\sigma_2^2) \qquad y = x_1 + x_2$$

$$p(y) = \mathcal{N}(y/\mu_1+\mu_2,\ \sigma_1^2+\sigma_2^2)$$

The Gaussian distribution can be generalised to more than one dimension

2.7. Multivariate models

Let us consider two random variables X and Y. The covariance is defined as

$$\mathrm{Cov}[X,Y] \triangleq \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$$

The covariance is a matrix of dimension D, where D is the dimension of the vector of the random variable. The diagonal is the variance. The correlation is defined as

$$\rho \triangleq \mathrm{corr}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}$$

<u>Remarks</u>:

- The fact that two variables are uncorrelated does not mean that they are independent. Example: $X \sim \mathrm{Unif}(-1,1)$ and $Y = X^2$, $\mathrm{corr}[X,Y]=0$. Conversely, two independent variables are uncorrelated.
- Correlation does not imply causality

The multivariate Gaussian distribution is defined as

$$\mathcal{N}(\vec{y}\,|\,\vec{\mu},\Sigma) \triangleq \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{y}-\vec{\mu})^T\Sigma^{-1}(\vec{y}-\vec{\mu})\right\}$$

$$\Sigma = \mathrm{Cov}[\vec{y}]$$

where D is the dimensionality of $\vec{y}$.