CSE 5523: Lecture Notes 2
Probability

Contents

2.1 Background: probability and probability spaces [Kolmogorov, 1933]

Probability is defined over a measure space $\langle O, \mathcal{E}, \mathsf{P} \rangle$ where the measure P (probability) sums to one.

This **probability measure space** $\langle O, \mathcal{E}, \mathsf{P} \rangle$ consists of:

1. a **sample space** O – a non-empty set of **outcomes**;
2. a **sigma-algebra** $\mathcal{E} \subseteq 2^O$ – a set of **events** which are subsets in the power set of O such that:
   (a) $\mathcal{E}$ contains O: $O \in \mathcal{E}$,
   (b) $\mathcal{E}$ is closed under complementation: $\forall_{A \in \mathcal{E}}\ O - A \in \mathcal{E}$,
   (c) $\mathcal{E}$ is closed under countable union: $\forall_{A_1..A_\infty \in \mathcal{E}} \bigcup_{i=1}^{\infty} A_i \in \mathcal{E}$

   (this set of events will serve as the domain of our probability function);
3. a **probability measure** $\mathsf{P} : \mathcal{E} \to \mathbb{R}_0^\infty$ – a function from events to non-negative reals such that:
   (a) the P measure is countably additive: $\forall_{A_1..A_\infty \in \mathcal{E} \text{ s.t. } \forall_{i,j}\ A_i \cap A_j = \emptyset}\ \mathsf{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathsf{P}(A_i)$,
   (b) the P measure of entire space is one: $\mathsf{P}(O) = 1$.

These are the **Kolmogorov axioms of probability**.

This characterization is helpful because it unifies probability spaces that may seem very different:

1. **discrete** spaces – e.g. a coin:

   $$\langle \underbrace{\{\text{H}, \text{T}\}}_{O}, \underbrace{\{\emptyset, \{\text{H}\}, \{\text{T}\}, \{\text{H}, \text{T}\}\}}_{\mathcal{E}}, \underbrace{\{\langle \emptyset, 0 \rangle, \langle \{\text{H}\}, .5 \rangle, \langle \{\text{T}\}, .5 \rangle, \langle \{\text{H}, \text{T}\}, 1 \rangle\}}_{\mathsf{P}} \rangle$$

2. **continuous** spaces – e.g. a dart (here $2^{\mathbb{R}^2}$ is a Borel algebra: a set of all open subsets of $\mathbb{R}^2$):

   $$\langle \underbrace{\mathbb{R}^2}_{O}, \underbrace{2^{\mathbb{R}^2}}_{\mathcal{E}}, \underbrace{\{\langle R, p \rangle \mid R \in 2^{\mathbb{R}^2}, p = \iint_{A \in R} \mathcal{N}_{\mathbf{0},\mathbf{1}}(x_A, y_A)\, dA\}}_{\mathsf{P}} \rangle$$

   (events must be open sets/ranges of outcomes because point outcomes have zero probability)

3. **joint** spaces using Cartesian products of sample spaces – e.g. two coins ($\{H,T\} \times \{H,T\}$):

$$\langle \underbrace{\{HH,HT,TH,TT\}}_{O}, \underbrace{\{\emptyset,\{HH\},\ldots,\{HH,HT,TH,TT\}\}}_{\mathcal{E}}, \underbrace{\{\langle\emptyset,0\rangle,\langle\{HH\},.25\rangle,\ldots,\langle\{HH,HT,TH,TT\},1\rangle\}}_{\mathsf{P}} \rangle$$

Also note: the set of outcomes can be larger than the set of events – e.g. a die used even/odd:

$$\langle \underbrace{\{1,2,3,4,5,6\}}_{O}, \underbrace{\{\emptyset,\{1,3,5\},\{2,4,6\},\{1,2,3,4,5,6\}\}}_{\mathcal{E}}, \underbrace{\{\langle\emptyset,0\rangle,\langle\{1,3,5\},.5\rangle,\langle\{2,4,6\},.5\rangle,\langle\{1,2,3,4,5,6\},1\rangle\}}_{\mathsf{P}} \rangle$$

This axiomatization entails, for any events (sets of outcomes) $A, B \in \mathcal{E}$:

1. $\mathsf{P}(A) \in \mathbb{R}_0^1$
2. $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$

Minimal events – those used as base cases in the closure operations – are called **atomic events**.

Atomic events in continuous models can have any size you want (like even/odd die), but not points.

Though probabilities are defined over sets of outcomes, we often write them using **propositions**.

For example, if $O = X \times Y$ and therefore $\forall_{o \in O}\ o = \langle x_o, y_o \rangle$:

$$\begin{array}{lll} \mathsf{P}(x) & = \mathsf{P}(X{=}x) & = \mathsf{P}(\{o \mid o{\in}O \wedge x_o{=}x\}) \quad \text{(allow any value for } y_o \text{ component)} \\ \mathsf{P}(x \wedge y) & = \mathsf{P}(X{=}x \wedge Y{=}y) & = \mathsf{P}(\{o \mid o{\in}O \wedge x_o{=}x \wedge y_o{=}y\}) \\ \mathsf{P}(\neg x) & = \mathsf{P}(X{\neq}x) & = \mathsf{P}(\{o \mid o{\in}O \wedge x_o{\neq}x\}) \end{array}$$

Random variables D are functions from outcomes x_o, y_o to **values**, e.g. distance of point to origin.

Often we will simply use Cartesian factors of a joint sample space (X, Y) as random variables.

Distributions are sometimes written as probabilities over (all values of) random variables:

$$\mathsf{P}(X) = \mathsf{P}(Y) \quad \Leftrightarrow \quad \forall_{x \in X} \forall_{y \in Y}\ \mathsf{P}(x) = \mathsf{P}(y).$$

We can also define **conditional probabilities** as ratios of these measures: $\mathsf{P}(y \mid x) = \frac{\mathsf{P}(x \wedge y)}{\mathsf{P}(x)}$.

2.2 A simple example

We can now distinguish some different kinds of (supervised) learning:

- **classification:** $\hat{y} = \mathrm{argmax}_y\ \mathsf{P}(y \mid x)$ with $y \in \mathbb{Z}^n$ (countable)
- **regression:** $\hat{y} = \mathrm{argmax}_y\ \mathsf{P}(y \mid x)$ with $y \in \mathbb{R}^n$ (uncountable)

We then define a **frequency space** $\langle O, \mathcal{E}, \mathsf{F} \rangle$ – same measure space with no $\mathsf{P}(O) = 1$ constraint.

We can define a frequency space using **counts** of some set of atomic events in some **training data**.

For example a model for fruits and colors:

$\langle$ {⟨apple,red⟩, ⟨apple,green⟩, ⟨pear,red⟩, ⟨pear,green⟩},
{∅, {⟨apple,red⟩}, {⟨apple,green⟩}, {⟨pear,red⟩}, {⟨pear,green⟩}, . . . }
{⟨∅, 0⟩, ⟨{⟨apple,red⟩}, 2⟩, ⟨{⟨apple,green⟩}, 1⟩, ⟨{⟨pear,red⟩}, 0⟩, ⟨{⟨pear,green⟩}, 2⟩, . . . } ⟩

(Counts for larger sets are simply sums, according to axiom 3a.)

We can now define a very simple machine learning example:

$$\mathsf{P}(A) = \frac{\mathsf{F}(A)}{\mathsf{F}(O)}$$

⟨ {⟨apple,red⟩, ⟨apple,green⟩, ⟨pear,red⟩, ⟨pear,green⟩},
{∅, {⟨apple,red⟩}, {⟨apple,green⟩}, {⟨pear,red⟩}, {⟨pear,green⟩}, . . . }
{⟨∅, 0⟩, ⟨{⟨apple,red⟩}, .4⟩, ⟨{⟨apple,green⟩}, .2⟩, ⟨{⟨pear,red⟩}, 0⟩, ⟨{⟨pear,green⟩}, .4⟩, . . . } ⟩

(Counts for larger sets are simply sums, according to axiom 3a.)

This is called **relative frequency estimation**.

2.3 Optimality of relative frequency estimation

Relative frequency estimation assigns the *highest* probability to your data!

Recall **combination** notation – number of orderings to choose $n_1, n_2, n_3, \ldots$ of each category:

$$\binom{\sum_j n_j}{n_1, n_2, n_3, \ldots} = \frac{(\sum_j n_j)!}{n_1!n_2!n_3!\ldots}$$

Using multinomial parameters $p_1, p_2, \ldots$, the probability of atomic event counts $n_1, n_2, \ldots$ is:

$$\binom{\sum_j n_j}{n_1, n_2, \ldots} \prod_j (p_j)^{n_j} = \binom{\sum_j n_j}{\bigtimes_j \{n_j\}} \prod_j (p_j)^{n_j}$$
$$= \binom{5}{2, 1, 0, 2} \mathsf{P}(\text{apple,red})^2 \, \mathsf{P}(\text{apple,green})^1 \, \mathsf{P}(\text{pear,red})^0 \, \mathsf{P}(\text{pear,green})^2$$

The parameters p_i that maximize probability of data are those where slope (derivative) is zero:

$$0 = \frac{\partial}{\partial p_i} \binom{\sum_j n_j}{\bigtimes_j \{n_j\}} \prod_j (p_j)^{n_j}$$

$$= \frac{\partial}{\partial p_i} \binom{\sum_j n_j}{n_i} (p_i)^{n_i} \binom{\sum_{j \neq i} n_j}{\bigtimes_{j \neq i} \{n_j\}} \prod_{j \neq i} (p_j)^{n_j} \qquad \text{definition of limit product}$$

$$= \frac{\partial}{\partial p_i} \binom{\sum_j n_j}{n_i} (p_i)^{n_i} (1 - p_i)^{\sum_{j \neq i} n_j} \qquad \text{multinomial distribution sums to one}$$

$$
\begin{aligned}
&= \binom{\sum_j n_j}{n_i} \frac{\partial}{\partial p_i} (p_i)^{n_i} (1-p_i)^{\sum_{j\neq i} n_j} && \text{product rule} \\
&= \frac{\partial}{\partial p_i} (p_i)^{n_i} (1-p_i)^{\sum_{j\neq i} n_j} && \text{division by } \binom{\sum_j n_j}{n_i} \\
&= \frac{\partial}{\partial p} p^n (1-p)^m && \text{let } p = p_i, n = n_i, m = \sum_{j\neq i} n_j \\
&= \left(\frac{\partial}{\partial p} p^n\right)(1-p)^m + p^n \left(\frac{\partial}{\partial p}(1-p)^m\right) && \text{product rule} \\
&= np^{n-1}(1-p)^m + p^n m (1-p)^{m-1} \left(\frac{\partial}{\partial p} 1 - p\right) && \text{power rule} \\
&= np^{n-1}(1-p)^m + p^n m (1-p)^{m-1}(-1) && \text{power rule} \\
&= p^{n-1}(1-p)^{m-1}(n(1-p) - mp) && \text{distributive axiom} \\
&= p^{n-1}(1-p)^{m-1}(n - np - mp) && \text{distributive axiom} \\
&= \underbrace{p^{n-1}}_{\text{root: } \hat{p}=0} \underbrace{(1-p)^{m-1}}_{\text{root: } \hat{p}=1} \underbrace{(n-(n+m)p)}_{\text{root: } \hat{p}=\frac{n}{n+m}} && \text{distributive axiom}
\end{aligned}
$$

So (ignoring the 0 and 1 roots, which are minima) the optimal parameters are all $\hat{p}_i = \frac{n_i}{\sum_j n_j}$.

This is called a **maximum likelihood estimate**.

2.4 Optimal continuous parameter estimation

'Normal' (Gaussian) distributions with parameters for mean μ and standard deviation σ:

$$
\mathcal{N}_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x-\mu)^2}{2\sigma^2}
$$

also have an easy optimal parameter estimate that maximizes the probability of data $x_1, x_2, \ldots$.

(If you are designing novel distributions, you may also want easy optimal parameter estimation!)

Again, the parameters μ, σ that maximize probability are those where slope (derivative) is zero:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \mu} \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x_i-\mu)^2}{2\sigma^2} \\
0 &= \frac{\partial}{\partial \mu} \ln \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x_i-\mu)^2}{2\sigma^2} && \text{max of function is max of log} \\
&= \frac{\partial}{\partial \mu} \sum_i \ln \left(\frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x_i-\mu)^2}{2\sigma^2}\right) && \text{log of product is sum of logs} \\
&= \sum_i \frac{\partial}{\partial \mu} \ln \left(\frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x_i-\mu)^2}{2\sigma^2}\right) && \text{sum rule} \\
&= \sum_i \frac{\partial}{\partial \mu} \ln \frac{1}{\sigma\sqrt{2\pi}} + \frac{-(x_i-\mu)^2}{2\sigma^2} && \text{log of product is sum of logs}
\end{aligned}
$$

$$= \sum_i \frac{\partial}{\partial \mu} \frac{-(x_i - \mu)^2}{2\sigma^2} \qquad \text{derivative of constant}$$

$$= \sum_i -\frac{1}{2\sigma^2} \frac{\partial}{\partial \mu}(x_i - \mu)^2 \qquad \text{product rule}$$

$$= \sum_i -\frac{1}{2\sigma^2}(-1)2(x_i - \mu) \qquad \text{power rule}$$

$$= \frac{1}{\sigma^2} \sum_i (x_i - \mu) \qquad \text{distributive axiom}$$

$$= \sum_i (x_i - \mu) \qquad \text{multiply by } \sigma^2$$

$$= \underbrace{-n\mu + \sum_i^n x_i}_{\text{root: } \hat{\mu} = \frac{1}{n}\sum_i^n x_i}$$

And for the standard deviation:

$$0 = \frac{\partial}{\partial \sigma} \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x_i - \mu)^2}{2\sigma^2}$$

$$0 = \frac{\partial}{\partial \sigma} \ln \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x_i - \mu)^2}{2\sigma^2} \qquad \text{max of function is max of log}$$

$$= \frac{\partial}{\partial \sigma} \sum_i \ln\left(\frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x_i - \mu)^2}{2\sigma^2}\right) \qquad \text{log of product is sum of logs}$$

$$= \sum_i \frac{\partial}{\partial \sigma} \ln\left(\frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(x_i - \mu)^2}{2\sigma^2}\right) \qquad \text{sum rule}$$

$$= \sum_i \frac{\partial}{\partial \sigma} \ln \frac{1}{\sigma\sqrt{2\pi}} + \frac{-(x_i - \mu)^2}{2\sigma^2} \qquad \text{log of product is sum of logs}$$

$$= \sum_i \frac{\partial}{\partial \sigma} - \ln(\sigma\sqrt{2\pi}) + \frac{-(x_i - \mu)^2}{2\sigma^2} \qquad \text{log of power}$$

$$= \sum_i \frac{\partial}{\partial \sigma} - \ln(\sqrt{\sigma^2 2\pi}) + \frac{-(x_i - \mu)^2}{2\sigma^2} \qquad \text{square root of square}$$

$$= \sum_i \left(\frac{\partial}{\partial \sigma} - \frac{1}{2}\ln(2\pi\sigma^2)\right) + \left(\frac{\partial}{\partial \sigma} \frac{-(x_i - \mu)^2}{2\sigma^2}\right) \qquad \text{sum rule}$$

$$= \left(\frac{\partial}{\partial \sigma} - \frac{n}{2}\ln(2\pi\sigma^2)\right) + \sum_i \left(\frac{\partial}{\partial \sigma} \frac{-(x_i - \mu)^2}{2\sigma^2}\right) \qquad \text{constant in discrete sum}$$

$$= -\frac{n}{2}\left(\frac{\partial}{\partial \sigma} \ln(2\pi\sigma^2)\right) + \sum_i \frac{1}{2}\left(-(x_i - \mu)^2 \frac{\partial}{\partial \sigma} \frac{1}{\sigma^2}\right) \qquad \text{product rule}$$

$$= -\frac{n}{2}\left(\frac{\partial}{\partial \sigma} \ln(2\pi\sigma^2)\right) + \sum_i \frac{1}{2}\left(-(x_i - \mu)^2 (-2)\frac{1}{\sigma^3}\right) \qquad \text{power rule}$$

$$= -\frac{n}{2}\left(\frac{\partial}{\partial\sigma}\ln(2\pi\sigma^2)\right) + \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2 \qquad \text{distributive axiom}$$

$$= -\frac{n}{2}\left(\frac{\partial}{\partial\sigma}\ln(2\pi) + \ln(\sigma^2)\right) + \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2 \qquad \text{log of product is sum of logs}$$

$$= -\frac{n}{2}\left(\frac{\partial}{\partial\sigma}\ln(\sigma^2)\right) + \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2 \qquad \text{derivative of constant}$$

$$= -\frac{n}{2}\left(\frac{\partial}{\partial\sigma}2\ln(\sigma)\right) + \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2 \qquad \text{log of power}$$

$$= -n\left(\frac{\partial}{\partial\sigma}\ln(\sigma)\right) + \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2 \qquad \text{product rule}$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_i (x_i-\mu)^2 \qquad \text{derivative of log}$$

$$= \underbrace{-n + \frac{1}{\sigma^2}\sum_i (x_i-\mu)^2}_{\text{root: } \hat{\sigma} = \sqrt{\frac{1}{n}\sum_i (x_i-\mu)^2}} \qquad \text{multiply by } \sigma$$

References

[Kolmogorov, 1933] Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin. Second English Edition, *Foundations of Probability* 1950, published by Chelsea, New York.