STAT 431 — Applied Bayesian Analysis — Course Notes

Probability Review

Spring 2019

Sample Space and Events

sample space: all possible outcomes (fully specified)

event: subset of sample space

Eg: Two coin flips

$$S \;=\; \text{sample space} \;=\; \{HH, HT, TH, TT\}$$
$$\text{event } A \;=\; \text{both flips same} \;=\; \{HH, TT\}$$

[Illustrate ...]

The usual set operations apply to events:

$$
\begin{aligned}
A \cup B &= \textbf{union} \text{ of } A \text{ and } B \\
&= \text{outcomes in either (or both)}
\end{aligned}
$$

$$
\begin{aligned}
A \cap B &= \textbf{intersection} \text{ of } A \text{ and } B \\
&= \text{outcomes in both}
\end{aligned}
$$

$$
\begin{aligned}
\overline{A} &= \textbf{complement} \text{ of } A \\
&= \text{outcomes not in } A
\end{aligned}
$$

Events A and B are **disjoint** if

$$A \cap B = \emptyset \quad \text{(the null set)}$$

Probability

probability: assigns to each event A a number $\mathrm{P}(A)$, with such properties as

- $0 \;\leq\; \mathrm{P}(A) \;\leq\; 1$
- $\mathrm{P}(\emptyset) \;=\; 0$ ($\emptyset$ is the "null event")
- $\mathrm{P}(S) \;=\; 1$ (S is the sample space)
- if A and B are disjoint,

$$\mathrm{P}(A \cup B) \;=\; \mathrm{P}(A) \,+\, \mathrm{P}(B)$$

- if $A_1, \ldots, A_n$ are disjoint,

$$\mathrm{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathrm{P}(A_i)$$

- $\mathrm{P}(\overline{A}) = 1 - \mathrm{P}(A)$

Eg: two fair coin flips

$$A = \text{both heads} \qquad B = \text{both tails}$$

$$\mathrm{P}(A) = ? \qquad \mathrm{P}(B) = ? \qquad \mathrm{P}\left(\overline{A}\right) = ?$$

$$\mathrm{P}(A \cup B) = ? \qquad \mathrm{P}(A \cap B) = ?$$

Conditioning

If $\mathrm{P}(B) \neq 0$, the **conditional probability** of A given B is

$$\mathrm{P}(A \mid B) \;=\; \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$$

($\mathrm{P}(A)$ is sometimes called the **marginal probability** of A)

Conditional probabilities behave like ordinary probabilities:

e.g. $0 \;\leq\; \mathrm{P}(A \mid B) \;\leq\; 1$

e.g. $\mathrm{P}\big(\overline{A} \mid B\big) \;=\; 1 \;-\; \mathrm{P}(A \mid B)$

Note:

$$\begin{aligned} \mathrm{P}(A \cap B) &= \mathrm{P}(B)\,\mathrm{P}(A \mid B) \qquad (\mathrm{P}(B) \neq 0) \\ &= \mathrm{P}(A)\,\mathrm{P}(B \mid A) \qquad (\mathrm{P}(A) \neq 0) \end{aligned}$$

In general,

$$\text{joint} = \text{marginal} \times \text{conditional}$$

Eg: two fair coin flips

$$A = \{HH, TT\} \qquad B = \{HH, HT, TH\}$$

$$\mathrm{P}(A) = ? \qquad \mathrm{P}(B) = ? \qquad \mathrm{P}(A \cap B) = ?$$

$$\mathrm{P}(A \mid B) = ?$$

$$\mathrm{P}(B \mid A) = ?$$

[Interpret ...]

Bayesian perspective:

$$\begin{aligned} M &= \text{a particular model for the data} \\ D &= \text{(event of) the data} \end{aligned}$$

$$\begin{aligned} \mathrm{P}(M) &= \text{probability of } M \text{ if we have no other information} \\ &= \text{“prior”} \end{aligned}$$

$$\begin{aligned} \mathrm{P}(D \mid M) &= \text{probability given to } D \text{ when } M \text{ is true} \\ &= \text{“likelihood”} \end{aligned}$$

$$\begin{aligned} \mathrm{P}(M \mid D) &= \text{probability of } M \text{ after observing } D \\ &= \text{“posterior”} \end{aligned}$$

Bayes' Rule

[Illustrate sample space ...]

Notice:

$$\mathrm{P}(A) \;\; = \;\; \mathrm{P}\big((A \cap B) \cup (A \cap \overline{B})\big)$$

Bayes' Rule

[Illustrate sample space ...]

Notice:

$$\begin{aligned} \mathrm{P}(A) &= \mathrm{P}\big((A \cap B) \cup (A \cap \overline{B})\big) \\ &= \mathrm{P}(A \cap B) + \mathrm{P}\big(A \cap \overline{B}\big) \end{aligned}$$

Bayes' Rule

[Illustrate sample space ...]

Notice:

$$\begin{aligned}
\mathrm{P}(A) &= \mathrm{P}\big((A \cap B) \cup (A \cap \overline{B})\big) \\
&= \mathrm{P}(A \cap B) \;+\; \mathrm{P}\big(A \cap \overline{B}\big) \\
&= \mathrm{P}(B)\ \mathrm{P}(A \mid B) \;+\; \mathrm{P}\big(\overline{B}\big)\ \mathrm{P}\big(A \mid \overline{B}\big)
\end{aligned}$$

(provided $0 < \mathrm{P}(B) < 1$)

Bayes' Rule (simple form):

$$\mathrm{P}(B \mid A) \;=\; \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}$$

Bayes' Rule (simple form):

$$\begin{aligned} \mathrm{P}(B \mid A) &= \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} \\ &= \frac{\mathrm{P}(B)\ \mathrm{P}(A \mid B)}{\mathrm{P}(B)\ \mathrm{P}(A \mid B)\ +\ \mathrm{P}\big(\overline{B}\big)\ \mathrm{P}\big(A \mid \overline{B}\big)} \end{aligned}$$

(provided all conditional probabilities exist)

Eg: Say the pop. of Cyprus is 80% Greek, 20% Turkish.

Suppose English is spoken by 90% of the Greeks and 50% of the Turks.

What's the prob. and English-speaking Cypriot is Greek?

$$A = \text{speaks English} \qquad\qquad B = \text{is Greek}$$

Expression for the desired probability?

$$\mathrm{P}(B) = ? \qquad\qquad \mathrm{P}\left(\overline{B}\right) = ?$$

$$\mathrm{P}(A \mid B) = ? \qquad\qquad \mathrm{P}\left(A \mid \overline{B}\right) = ?$$

Answer?

Now generalize ...

Suppose $B_1, B_2, B_3, \ldots$ form a **partition** of S:

- all are disjoint
- $\bigcup_{\text{all } j} B_j \ = \ S$ (exhaustive)

Also, assume $\mathrm{P}(B_j) \neq 0$, all j.

Law of Total Probability:

$$\mathrm{P}(A) \ = \ \sum_{\text{all } j} \mathrm{P}(B_j)\ \mathrm{P}(A \mid B_j)$$

[Illustrate ...]

Bayes' Rule (for probabilities):

If $B_1, B_2, \ldots$ is a partition,

$$\mathrm{P}(B_i \mid A) \quad = \quad \frac{\mathrm{P}(B_i)\ \mathrm{P}(A \mid B_i)}{\sum_{\text{all } j} \mathrm{P}(B_j)\ \mathrm{P}(A \mid B_j)}$$

(The previous special case had $B_1 = B$, $B_2 = \overline{B}$.)

Bayes' Rule (for probabilities):

If $B_1, B_2, \ldots$ is a partition,

$$\mathrm{P}(B_i \mid A) \quad = \quad \frac{\mathrm{P}(B_i)\ \mathrm{P}(A \mid B_i)}{\sum_{\text{all } j} \mathrm{P}(B_j)\ \mathrm{P}(A \mid B_j)}$$

(The previous special case had $B_1 = B$, $B_2 = \overline{B}$.)

So

$$\mathrm{P}(B_i \mid A) \quad \propto \quad \mathrm{P}(B_i)\ \mathrm{P}(A \mid B_i)$$

(since the denominator doesn't depend on i)

Bayesian application:

$$\begin{aligned} M_1, M_2, \ldots &= \text{distinct models for the data} \\ D &= \text{(event of) the data} \end{aligned}$$

By Bayes' Rule,

$$\begin{aligned} \mathrm{P}(M_i \mid D) &\propto \mathrm{P}(M_i)\, \mathrm{P}(D \mid M_i) \\ \text{posterior} &\propto \text{prior} \times \text{likelihood} \end{aligned}$$

The (inverse) proportionality constant

$$\sum_{\text{all } j} \mathrm{P}(M_j)\, \mathrm{P}(D \mid M_j)$$

is called the **normalizing constant**.

Eg: Waldo (revisited)

Independent Events

Events A and B are **independent** if

$$\mathrm{P}(A \cap B) \;=\; \mathrm{P}(A)\,\mathrm{P}(B)$$

(otherwise **dependent**)

If $\mathrm{P}(B) \neq 0$, this is the same as

$$\mathrm{P}(A \mid B) \;=\; \mathrm{P}(A)$$

$$(\text{conditional} \;=\; \text{marginal})$$

Events A and B are **conditionally independent** given C if

$$\mathrm{P}(A \cap B \mid C) \;\; = \;\; \mathrm{P}(A \mid C)\, \mathrm{P}(B \mid C)$$

Note: A and B are not necessarily independent!

Often there are data events that are independent conditional on the model:

$$\mathrm{P}(D_1 \cap D_2 \mid M) \;\; = \;\; \mathrm{P}(D_1 \mid M)\, \mathrm{P}(D_2 \mid M)$$

That is, the likelihood may factor.

Random Variables and Distributions

random variable: real-valued function on the sample space

May be ...

- **discrete**: takes values in a countable set

  e.g. binomial, geometric, Poisson

- **continuous**: takes values on a continuum

  e.g. normal, exponential, gamma

The **distribution** of a random variable X is characterized by its **density**:

- Discrete density:

$$p(x) \;\; = \;\; \mathrm{P}(X = x)$$

(sometimes called a "mass function")

- Continuous density: $\quad p(x)$ such that

$$\int_G p(x)\, dx \;\; = \;\; \mathrm{P}(X \in G)$$

(often called a p.d.f.)

The **joint distribution** of random variables X and Y can often be characterized by a **joint density**

$$p(x, y)$$

- Both discrete:

$$p(x, y) \;=\; \mathrm{P}(X = x, Y = y) \;=\; \mathrm{P}\big(\{X = x\} \cap \{Y = y\}\big)$$

- Jointly continuous:

$$\int_{G_1} \int_{G_2} p(x, y)\, dy\, dx \;=\; \mathrm{P}(X \in G_1, Y \in G_2)$$

The individual densities of X and Y are their **marginal densities**, which define their **marginal distributions**.

Eg:

$$p(x) = \begin{cases} \displaystyle\sum_{\text{all } y} p(x,y), & Y \text{ discrete} \\ \displaystyle\int p(x,y)\,dy, & Y \text{ continuous} \end{cases}$$

Conditioning

The **conditional distribution** of X given Y is characterized by the **conditional density**

$$p(x \mid y) \;=\; \frac{p(x, y)}{p(y)} \qquad \text{(wherever } p(y) > 0\text{)}$$

Conditioning

The **conditional distribution** of X given Y is characterized by the **conditional density**

$$p(x \mid y) \;=\; \frac{p(x,y)}{p(y)} \qquad \text{(wherever } p(y) > 0\text{)}$$

Note:

$$p(x,y) \;=\; p(y)\,p(x \mid y) \;=\; p(x)\,p(y \mid x)$$

is another example of the general form

$$\text{joint} \;=\; \text{marginal} \times \text{conditional}$$

This idea can be used to define the joint density when X and Y are of different types.

For example, if X is continuous and Y is discrete, let

$$p(x, y) \;=\; p(y)\, p(x \mid y) \;=\; p(x)\, p(y \mid x)$$

where

$$\begin{aligned} p(x \mid y) &= \text{a continuous density for each } y \\ p(y \mid x) &= \text{a discrete density for each } x \end{aligned}$$

(Use whichever of these is most convenient.)

A general process for working with the joint distribution of X and Y:

1. Specify the marginal density of X
2. Specify the conditional density of Y given X
3. Use the product of these densities as their joint density

Example: Uniform-Binomial

$$\begin{aligned} X &\sim \text{uniform}(0, 1) \\ Y \mid X = x &\sim \text{binomial}(n, x) \end{aligned}$$

(n is a given "number of trials", x is "success prob.")

Example: Uniform-Binomial

$$\begin{aligned} X &\sim \text{uniform}(0,1) \\ Y \mid X = x &\sim \text{binomial}(n, x) \end{aligned}$$

(n is a given "number of trials", x is "success prob.")

So

$$p(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$p(y \mid x) = \binom{n}{y} x^y (1-x)^{n-y} \qquad y = 0, \ldots, n$$

... and the "joint density" is

$$p(x)\,p(y \mid x) \;=\; \begin{cases} \dbinom{n}{y} x^y\,(1-x)^{n-y} & 0 < x < 1,\ y = 0, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

The marginal density for X is (of course)

$$p(x) \;=\; \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

The marginal density for Y is (for $y = 0, \ldots, n$)

$$
\begin{aligned}
p(y) &= \int p(x)\, p(y \mid x)\, dx \\
&= \int_0^1 \binom{n}{y} x^y (1-x)^{n-y}\, dx \\
&= \binom{n}{y} \int_0^1 \underbrace{x^y (1-x)^{n-y}}_{\text{``kernel'' of a beta density}}\, dx
\end{aligned}
$$

The marginal density for Y is (for $y = 0, \ldots, n$)

$$
\begin{aligned}
p(y) &= \int p(x)\, p(y \mid x)\, dx \\
&= \int_0^1 \binom{n}{y} x^y (1-x)^{n-y}\, dx \\
&= \binom{n}{y} \int_0^1 \underbrace{x^y (1-x)^{n-y}}_{\text{“kernel” of a beta density}}\, dx
\end{aligned}
$$

Recall density of $\mathrm{beta}(\alpha, \beta)$ distribution:

$$
\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1} (1-x)^{\beta-1} \qquad 0 < x < 1
$$

(see Cowles, Table A.2)

Thus, for $y = 0, \ldots, n$,

$$
\begin{aligned}
p(y) &= \binom{n}{y} \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)} \\
&\quad \cdot \int_0^1 \underbrace{\frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\, x^y\, (1-x)^{n-y}}_{\mathrm{beta}(y+1,\, n-y+1) \text{ density}}\, dx
\end{aligned}
$$

Thus, for $y = 0, \dots, n$,

$$p(y) = \binom{n}{y} \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$

$$\cdot \int_0^1 \underbrace{\frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} x^y (1-x)^{n-y}}_{\mathrm{beta}(y+1,\, n-y+1) \text{ density}} \, dx$$

$$= \binom{n}{y} \frac{y!\,(n-y)!}{(n+1)!} \cdot 1$$

Thus, for $y = 0, \ldots, n$,

$$
\begin{aligned}
p(y) &= \binom{n}{y} \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)} \\
&\quad \cdot \int_0^1 \underbrace{\frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\, x^y\,(1-x)^{n-y}}_{\mathrm{beta}(y+1,\,n-y+1)\text{ density}}\; dx \\
&= \binom{n}{y} \frac{y!\,(n-y)!}{(n+1)!} \quad \cdot \quad 1 \\
&= \frac{n!}{y!\,(n-y)!}\,\frac{y!\,(n-y)!}{(n+1)\cdot n!} \quad = \quad \frac{1}{n+1}
\end{aligned}
$$

So the marginal distribution of Y is a "discrete uniform" distribution:

$$p(y) \quad = \quad \begin{cases} \dfrac{1}{n+1} & y = 0, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

[Illustrate ...]

Bayes' Rule

Bayes' Rule (for densities):

$$p(y \mid x) \;=\; \frac{p(y)\, p(x \mid y)}{C}$$

where

$$C \;=\; p(x) \;=\; \begin{cases} \displaystyle\sum_{\text{all } y} p(y)\, p(x \mid y), & Y \text{ discrete} \\[2ex] \displaystyle\int p(y)\, p(x \mid y)\, dy, & Y \text{ continuous} \end{cases}$$

C is the **normalizing constant**.

Bayesian application:

Suppose the model is (fully) defined by a **parameter** θ.

Let y be the observed data.

Then

$$\begin{array}{ccccc} p(\theta \mid y) & \propto & p(\theta) & \cdot & p(y \mid \theta) \\ \text{posterior} & \propto & \text{prior} & \times & \text{likelihood} \end{array}$$

where the proportionality is in θ (for fixed y).

(The likelihood is sometimes written as $L(\theta; y)$.)

More Variables

Densities can be extended to three or more variables.

E.g. X, Y, and Z could have a joint density defined by

$$p(x, y, z) \;=\; p(x)\, p(y \mid x)\, p(z \mid x, y)$$

where conditioning on two variables is defined as, e.g.

$$p(z \mid x, y) \;=\; \frac{p(x, y, z)}{p(x, y)} \qquad \text{(wherever } p(x, y) > 0\text{)}$$

The marginal densities would be denoted

$$p(x, y), \qquad p(x, z), \qquad p(y, z),$$
$$p(x), \qquad p(y), \qquad p(z)$$

Marginal densities are obtained by summing/integrating out the other variables, e.g.

$$p(x, z) = \begin{cases} \displaystyle\sum_{\text{all } y} p(x, y, z), & Y \text{ discrete} \\ \displaystyle\int p(x, y, z)\, dy, & Y \text{ continuous} \end{cases}$$

Similarly, joint conditionals can be defined as, e.g.

$$p(x, y \mid z) \;\; = \;\; \frac{p(x, y, z)}{p(z)} \qquad \text{(wherever } p(z) > 0\text{)}$$

Certain rules for marginal densities extend to conditional densities, e.g.

$$p(x, y \mid z) \;\; = \;\; p(x \mid z)\, p(y \mid x, z)$$

Independent Random Variables

X and Y are **independent** when they have a joint density that factors into marginals:

$$p(x, y) \;=\; p(x)\, p(y)$$

Note: If X and Y are independent,

$$p(x \mid y) \;=\; p(x) \qquad\qquad p(y \mid x) \;=\; p(y)$$

Note: If $p(x \mid y)$ doesn't depend on y (or if $p(y \mid x)$ doesn't depend on x), then X and Y are independent. (Why?)

Let Z be another random variable.

X and Y are **conditionally independent given** $Z = z$ if

$$p(x, y \mid z) \;\; = \;\; p(x \mid z)\, p(y \mid z)$$

In general, this does not imply that X and Y are (marginally) independent.

X and Y are **conditionally independent given** Z if

$$p(x, y \mid z) \;=\; p(x \mid z)\, p(y \mid z) \qquad \text{for all } z \quad (p(z) > 0)$$

This is (almost) equivalent to

$$p(x \mid y, z) \;=\; p(x \mid z)$$

and to

$$p(y \mid x, z) \;=\; p(y \mid z)$$

Measures of Location and Spread

The **expected value** or **mean** of X is

$$\mathrm{E}(X) = \begin{cases} \displaystyle\sum_{\text{all } x} x\, p(x), & X \text{ discrete} \\ \displaystyle\int x\, p(x)\, dx, & X \text{ continuous} \end{cases}$$

A **median** m_X of X satisfies

$$\mathrm{P}(X < m_X) \leq 0.5 \quad \text{and} \quad \mathrm{P}(X > m_X) \leq 0.5$$

A **mode** of X is a value maximizing $p(x)$. It need not exist or be unique.

An α-**quantile** x_α of X satisfies

$$\mathrm{P}(X < x_\alpha) \;\; \leq \;\; \alpha \qquad \text{and} \qquad \mathrm{P}(X > x_\alpha) \;\; \leq \;\; 1 - \alpha$$

If X is continuous,

$$\mathrm{P}(X \leq x_\alpha) \;\; = \;\; \alpha$$

[Illustrate ...]

The **variance** of X is

$$\mathrm{Var}(X) \;=\; \mathrm{E}\big((X-\mu_X)^2\big)$$

where $\mu_X = \mathrm{E}(X)$.

An **interquartile range (IQR)** of X is

$$x_{0.75} \;-\; x_{0.25}$$

(i.e. the difference between the first and third quartile)

The **conditional expected value** (or **conditional mean**) of X given $Y = y$ is

$$\mathrm{E}(X \mid Y = y) \;=\; \begin{cases} \displaystyle\sum_{\text{all } x} x\, p(x \mid y), & X \text{ discrete} \\[2ex] \displaystyle\int x\, p(x \mid y)\, dx, & X \text{ continuous} \end{cases}$$

The **conditional variance** of X given $Y = y$ is

$$\mathrm{Var}(X \mid Y = y) \;=\; \mathrm{E}\big((X - \mu_{X|y})^2 \mid Y = y\big)$$

where $\mu_{X|y} \;=\; \mathrm{E}(X \mid Y = y)$.

Notational note:

We sometimes write

$$\mathrm{E}(X \mid y) \quad \text{for} \quad \mathrm{E}(X \mid Y = y)$$

$$\mathrm{Var}(X \mid y) \quad \text{for} \quad \mathrm{Var}(X \mid Y = y)$$

Similarly, write

$$X \mid y \;\sim\; \cdots \qquad \text{for} \qquad X \mid Y = y \;\sim\; \cdots$$

Transformation of Variables

Suppose X is continuous, with density $p(x)$, and let

$$Y = g(X)$$

where g has a differentiable inverse g^{-1}.

Then Y is continuous, with density

$$p(y) = p(x) \left| \frac{dx}{dy} \right|$$

where x is (implicitly) equal to $g^{-1}(y)$.

This **transformation-of-variables formula** is sometimes more explicitly written as

$$p_Y(y) \quad = \quad p_X\big(g^{-1}(y)\big) \left| \frac{d}{dy}\, g^{-1}(y) \right|$$

where p_X and p_Y are densities of X and Y.