pdf's, cdf's, conditional probability

September 17, 2013

Densities

- In $\mathbb{R}^n$ any function $p : \mathbb{R}^n \to \mathbb{R}$ satisfying $p(x) \geq 0$ for all $x \in \mathbb{R}^n$ and $\int_{\mathbb{R}^n} p(x)dx = 1$ can be used to define probabilities of sets in $\mathbb{R}^n$ and expectations of functions on $\mathbb{R}^n$. The function p is then called the **density**, or **pdf** (for probability density function) for the probability it defines.

- As a reminder:

$$P(A) = \int_{x \in A} p(x)\, dx \;, \qquad E[f] = \int_{\mathbb{R}^n} f(x)p(x)\, dx \;.$$

Densities for distributions with discontinuities

The only kind of discontinuous distributions we will be concerned with are ones that lump together continuous distributions over sets of different dimensions. That is we consider cases where

- $S = S_1 \cup \cdots \cup S_n$ with the S_j disjoint;
- Each S_j is a subset of $\mathbb{R}^{n_j}$ for some n_j, embedded in $S = \mathbb{R}^n$; (Technically, each S_j maps isometrically into a subset of $\mathbb{R}^{n_j}$.)

- For $A_j \subset S_j$,

$$P(A_j) = \int_{A_j} p(x(z))\,dz \ ,$$

where the integral is interpreted as an ordinary integral w.r.t. $z \in \mathbb{R}^{n_j}$, $x(z)$ maps points in $\mathbb{R}^{n_j}$ into the corresponding points in $\mathbb{R}^n$, and $p(x)$ is what we define as the density function for this distribution, over all of $\mathbb{R}^n$. Special case: $S_j = \mathbb{R}^0$, i.e. $S_j = \{x_j\}$, a single point. Then the "integral" is just $p(x_j)$.

- Then we can find the probability of any $A \subset \mathbb{R}^n$, in the σ-field $\mathcal{B}$ generated by the rectangles, from

$$P(A) = \sum_{j=1}^{n} P(A \cap S_j) \ .$$

- Morals of the story: There is *always* a density function, and we can always break up calculations of probability and expectations into pieces, all of which involve just ordinary integration. (There are more complicated situations, which we won't encounter in this course, where the part about ordinary integration isn't true.)

- Examples: Catalog prices (concentration on \$9.99, \$19.99, etc.); Gasoline purchase amounts (concentration on \$10/price, \$20/price, etc.); minimum wage example (concentration on $w = w_{\text{min}}$).

cdf's

- The cdf (cumulative distribution function) of the n-dimensional random vector X is defined by

$$F_X(a) = P[X \leq a] = P[X_i \leq a_i,\ i = 1, \ldots, n]\,.$$

- Useful to plot, easy to characterize in $\mathbb{R}^1$. F is a cdf for a univariate random variable if and only if $F(x) \to 0$ as $x \to -\infty$, $F(x) \to 1$ as $x \to \infty$, and F is monotonically increasing. $P[(a, b]] = F(b) - F(a)$.

- In 2d, it is not true that any monotonically increasing function that tends to 0 at $-\infty$ and to 1 at $+\infty$ is a cdf.

- Additional necessary condition in $\mathbb{R}^n$ is that F imply that all rectangles

$$\{x \mid a_1 < x_1 \leq b_1, \ldots, a_n < x_n \leq b_n\} = r(a,b)$$

have positive probability. This translates in 2d to

$$F(b_1, b_2) + F(a_1, a_2) - F(a_1, b_2) - F(b_1, a_2) \geq 0\,, \quad \text{all } a \leq b \in \mathbb{R}^n\,.$$

- Expressing probabilities of rectangles with cdf values becomes more and more messy as n increases.

- Sufficient conditions, in addition to the 0 and 1 limits, that an n times differentiable function F on $\mathbb{R}^n$ be a cdf: $\partial^n F / \partial x_1 \ldots \partial x_n \geq 0$ everywhere, in which case this partial derivative is the density function.

- cdf's are widely used to characterize and analyze one-dimensional distributions. Higher dimensional cdf's don't turn up often in applied work.

Conditional expectation

Suppose we have a random variable Y and a random vector X, defined on the same probability space S.

- The **conditional expectation** of Y given X is written as $E[Y \mid X]$.
- It is a function of X alone.
- For any continuous, bounded function g of X, $E[g(X)Y] = E\big[g(X)E[Y \mid X]\big]$.
- This property *defines* conditional expectation.
- Conditional expectation is unique, except that if $f(X)$ and $h(X)$ both satisfy the defining property for $E[Y \mid X]$, it is possible that $f(X) \neq h(X)$ on a set of X values of probability zero.

Special optional slide for anyone who knows measure theory and doubts that C.E.'s always exist

- For any random variable Y with finite expectation, we can define, by $\sigma_Y(A) = E[\mathbf{1}_A \cdot Y]$, a set function on the σ-field $\mathcal{B}_{Y,X}$ generated by rectangles in Y, X-space.

- σ_Y is continuous w.r.t. the joint probability measure on Y, X space — that is, if $P[A] = 0$, then $\sigma_Y(A) = 0$. This is clear because σ_Y is by construction a set function whose Radon-Nikodym derivative w.r.t. probability on X, Y space is Y.

- If we restrict σ_Y to $\mathcal{B}_X$, the sub-σ-field generated by X, is of course still absolutely continuous and has a Radon-Nikodym derivative w.r.t.

P restricted to this sub-σ-field. The only regularity condition necessary for this is that σ_Y restricted to $\mathcal{B}_X$ is σ-finite, and since Y has finite expectation, this is automatic.

- The Radon-Nikodym derivative of σ_Y w.r.t. P restricted to $\mathcal{B}_X$ is $E[Y \mid X]$.

Conditional probability

- $P[A \mid X] = E[\mathbf{1}_A \mid X]$.

- Interesting question: Is $P[\,\cdot \mid X]$ defined this way a well-behaved probability function on $\mathcal{B}$ for every X, or at least for a set of X's with probability 1? Too interesting for us. The answer is yes for the situations we will encounter in this course.

- In the standard purely purely continuous case, there is a **conditional pdf**, which can be found from the formula

$$p(y \mid x) = \frac{p(y, x)}{\int p(y, x)\, dy}\,.$$

- In the pure discrete case (y_i and x_j each take on only finitely many values on S) the corresponding formula is

$$p(y_i \mid x_j) = \frac{p(y_i, x_j)}{\sum_k p(y_k, x_i)} \,.$$

- The discrete formula is a special case of the continuous one if we use Lebesgue integration in the denominator and use the natural interpretation of what the S_j's are for the integral. In the simplest mixed discrete-continuous cases, where the S_j's are all isolated points except for one, say S_1, that is the rest of S, the integral formula also applies, again with the natural interpretation of what the S_j's are when we integrate w.r.t y.

- In more complicated situations, though, where the S_j's have positive dimension, the simple density-based formula cannot be relied on. This occurs rarely, so we will not attempt to discuss general rules for generated conditional densities in this case. If you encounter it in research (or, maybe, in a problem set), you can handle it by going back to the defining property of conditional expectation.

Marginal distributions

If X and Y are two random vectors defined on the same probability space and with joint density $p(x, y)$, the **marginal pdf** of X is $\pi(x) = \int p(x, y)\, dy$. It can be used to determine the probability of any set A defined entirely in terms X, i.e.

$$P[A] = \int_A p(x, y)\, dx\, dy = \int_A \left(\int p(x, y)\, dy \right) dx = \int_A \pi(x)\, dx\ .$$

The second equality follows because the restriction of the domain of integration to A puts no constraint on y, because by assumption A is defined entirely in terms of x.

With this definition, we can see that the rule for forming a conditional density from a joint density can also be written more compactly as $p(y \mid x) = p(x, y)/\pi(x)$

Inverse probability and Bayes' rule

- A common situation: There is a "parameter" β whose value we don't know, but we believe that a random variable Y has a distribution, conditional on β, with density $p(y \mid \beta)$.

- Before we observe Y our uncertainty about β is characterized by the pdf $\pi(\beta)$.

- The rule for forming conditional densities from joint can be solved to give us the joint pdf of y and β: $q(y, \beta) = p(y \mid \beta)\pi(\beta)$.

- Applying the rule again, we get the conditional pdf of $\{\beta \mid Y\}$ as

$$r(\beta \mid y) = \frac{p(y \mid \beta)\pi(\beta)}{\int p(y \mid \beta)\pi(\beta)\, d\beta}\,.$$

- This is **Bayes' rule**.

Independence

- If two random vectors X and Y have joint pdf $p(x, y)$, they are **independent** if and only if $p(x, y) = q_X(x)q_Y(y)$, where q_X and q_Y both integrate to one.

- In this case it is easy to verify that q_X and q_Y are the marginal pdf's of X and Y and also $q_X(x) = q_{X|Y}(x|y)$, $q_Y(y) = q_{Y|X}(y|x)$, that is, q_X and q_Y are also the conditional pdf's of $X \mid Y$ and $Y \mid X$.

- Obviously this means that the conditional distribution of $\{Y|X\}$ does not depend on X and for any function f of Y, $E[f(Y) \mid X] = E[f(Y)]$. (Of course also the same things with the Y, X roles reversed.)

- A more general definition: Y is independent of X if for every function $g(Y)$ such that $E[|g(Y)|] < \infty$, $E[g(Y) \mid X] \equiv E[g(Y)]$. It turns out that if this is true, the same is true with the roles of x and y reversed.

- A collection $\{X_1, \ldots, X_n\}$ of random vectors is **mutually independent** if for every i and for every g with $E[g(X_i)]$ defined and finite, $E[g(X_i) \mid X_{-i}] = E[g(X_i)]$. Here we're using the notation that X_{-i} means all the elements of the X vector except the one with index i. If they have a joint pdf, this is equivalent to

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} q_i(x_i) \ .$$

- It is possible to have X_i independent of X_j for any $i \neq j$ between 1 and n, yet to have the collection $\{X_1, \ldots, X_n\}$ not mutually independent. That is, pairwise independence does not imply mutual independence.