COMS30035, Machine learning: Probabilistic Graphical Models 0

James Cussens
james.cussens@bristol.ac.uk

Department of Computer Science, SCEEM
University of Bristol

October 9, 2020

Random variables

- Virtually all machine learning / statistics is done using *random variables*.
- Here's a simple random variable (r.v.) for modelling dice-throwing: Dice is a r.v. whose *domain* is $\{1, 2, 3, 4, 5, 6\}$ and we have a probability distribution over the possible values of Dice.
- One possible distribution is $P(\mathrm{Dice} = x) = 1/6$ for all $x \in \{1, 2, 3, 4, 5, 6\}$

Continuous random variables

- In the case of discrete random variables we can define its probability distribution by simply tabulating the probabilities.
- This is evidently not possible for a continuous random variable (i.e. tomorrow's temperature at noon).
- Instead for continuous random variables we have a *probability density function*.

Gaussian distribution

Here is the probability density function (p.d.f.) for a Gaussian distribution with mean μ and variance σ^2:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

Density functions

- Instead of requiring that a finite set of probabilities add up to 1, if f is a probability density function then the area under its curve must equal 1.

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \tag{2}$$

- Also $f(x) \geq 0$ for all x.

Using density functions

- We can use the pdf to get the probability that some random variable X takes a value in any given interval

$$P(X \in [a, b]) = \int_{x=a}^{x=b} f(x)\, dx \tag{3}$$

- In general this integral need not be easy to compute.
- If X is a random variable with a Gaussian distribution what is $P(X = 0)$?

Means and modes

Mean (or *expected value*):

$$E(X) = \sum_x x \; P(X = x)$$

$$E(X) = \int_{x=-\infty}^{x=\infty} x \; f(x) \; dx$$

Mode

$$\arg\max_x P(X = x)$$

$$\arg\max_x f(x)$$

Variance

- Let the mean of some random variable X be μ.
- Consider the function $g(x) = (\mu - x)^2$ which measures squared difference from the mean.
- $g(X)$ has a distribution (it's a function of a random variable) so it too has a mean.
- This value is called the *variance* of X; it is the expected squared distance from the mean and so measures the spread of the distribution.

$$\mathrm{Var}(X) = \sigma^2(X) = \sum_x P(X = x)(\mu - x)^2$$

$$\mathrm{Var}(X) = \sigma^2(X) = \int_{x=-\infty}^{x=\infty} f(x)(\mu - x)^2$$

- The square root of the variance is called the *standard deviation*, denoted $\sigma(X)$

Multivariate distributions

- A *multivariate distribution* is one defined using two or more random variables. The term *joint distribution* is also often used.
- For example, to model the outcome of throwing two dice we could have two random variables D_1 and D_2.
- To define the joint distribution in this case we need to specify a value for every combination of values (*joint instantiation*) for these two random variables.
- For example, one joint distribution is:

$$P(D_1 = x, D_2 = y) = \frac{1}{36} \quad \forall x, y \in \{1, 2, 3, 4, 5, 6\}$$

Marginal distributions

- From a joint distribution we can produce *marginal distributions* over any subset of the random variables, by ‘summing out’ (aka ‘marginalising away’) the random variables we don’t want.
- For example, from the following joint distribution $P(X, Y)$ over two binary random variables X and Y, we can produce two marginal distributions: $P(X)$ (in the bottom ‘margin’) and $P(Y)$ (in the ‘margin’ on the right).

$P(X, Y)$	$X = 0$	$X = 1$	$P(Y)$
$Y = 0$	0.2	0.3	0.5
$Y = 1$	0.4	0.1	0.5
$P(X)$	0.6	0.4	

Marginal distributions (ctd)

- The process of producing a marginal distribution is known as *marginalisation*.
- You should think of marginalisation as projecting a higher-dimensional distribution to get a lower dimensional one.
- Here is how we 'marginalise out' X_1 from a k-dimensional joint distribution over the variables $X_1, \ldots, X_k$.

$$P(X_2 = x_2, \ldots X_k = x_k) = \sum_{x_1} P(X_1 = x_1, X_2 = x_2, \ldots X_k = x_k)$$

Independence

- Two discrete random variables X and Y are *independent* if (and only if)

$$P(X = x, Y = y) = P(X = x)P(Y = y) \quad \forall x, y \tag{4}$$

- Are X and Y independent in the following distribution?

$P(X, Y)$	$X = 0$	$X = 1$	$P(Y)$
$Y = 0$	0.2	0.3	0.5
$Y = 1$	0.4	0.1	0.5
$P(X)$	0.6	0.4	

Conditional distributions

- Let X and Y be two discrete distributions, then the distribution over X *conditional on* $Y = y$ or *given* $Y = y$ is:

$$P(X|Y = y) = \frac{P(X, Y = y)}{P(Y = y)} \tag{5}$$

- Note: $P(X|Y = y)$ is undefined if $P(Y = y) = 0$ (that makes sense, no?)
- Conditional distributions are the cornerstone of statistics/machine learning, since we condition on the observed data to get distributions over unknown quantities.
- In the Bayesian approach to statistics/machine learning that's pretty much all we do!

Bayes theorem

- Since $P(X, Y) = P(X)P(Y|X) = P(Y)P(X|Y)$ we can re-arrange to get *Bayes theorem*

$$P(Y|X) = \frac{P(Y)P(X|Y)}{P(X)} \tag{6}$$

Suppose θ were some parameter and we observed some data $D = d$, then Bayes theorem tells us that:

$$P(\theta|D = d) = \frac{P(\theta)P(D = d|\theta)}{P(D = d)} \tag{7}$$

- $P(\theta)$ is the *prior distribution* for θ.
- $P(D|\theta)$ is known as the *likelihood*.

Continuous multivariate distributions

- So far I have defined: joint distributions, marginal distributions, conditional distributions, and independence all in terms of discrete distributions.
- But all these concepts apply (of course!) to continuous multivariate distributions.
- Everything is pretty much the same except addition is replaced by integration and a finite set of probabilities is replaced by probability density functions.

Continuous joint distributions and marginals

- A joint continuous distribution over, say, two variables X and Y is defined by a a probability density function with two arguments.
- Suppose this pdf was denoted $f_{X,Y}$ then here's how to get the marginal over just X:

$$f_X(x) = \int_y f_{X,Y}(x,y)dy \tag{8}$$

Continuous joint distributions and conditioning

- Given $f_{X,Y}$ we can define a conditional distribution by simple division:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

- For the conditional distribution to be defined we need $f_X(x) > 0$

Independence

- For two continuous random variables X and Y to be independent we must have:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \tag{9}$$

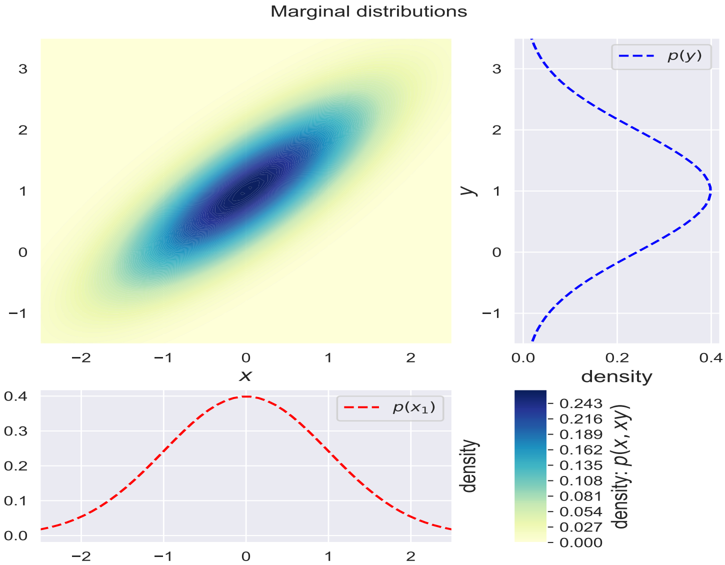
Multivariate Gaussian distribution

The most important multivariate distribution is the multivariate Gaussian distribution. Here's the p.d.f for a k-dimensional Gaussian distribution:

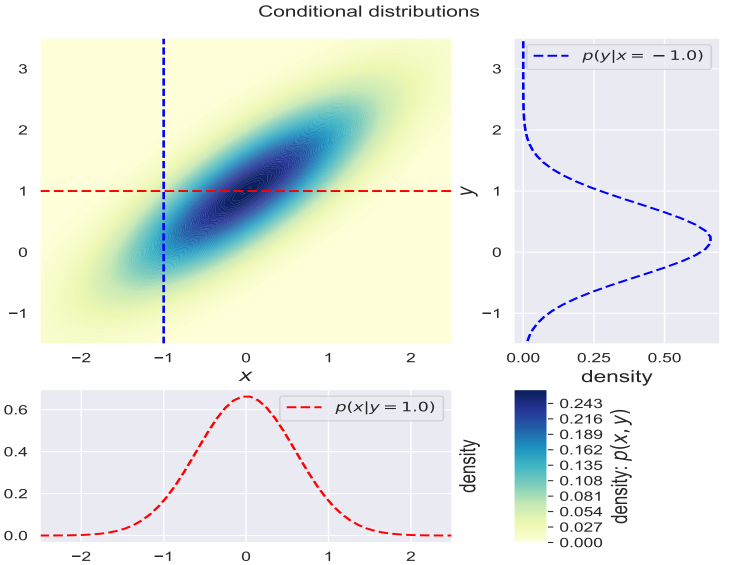
$$f(x_1, \ldots, x_k) = f(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)}{\sqrt{(2\pi)^k |\mathbf{\Sigma}|}} \tag{10}$$

- Instead of a single number (a scalar) as a mean, we now have a k-dimensional mean *vector* μ.
- Instead of a scalar variance, we now have a $k \times k$ *covariance matrix* $\mathbf{\Sigma}$.
- Let's plot some Gaussian density functions when $k = 2$.
- I used this Jupyter notebook written by Peter Roelants (ML Engineer at Twitter) to produce the following plots.

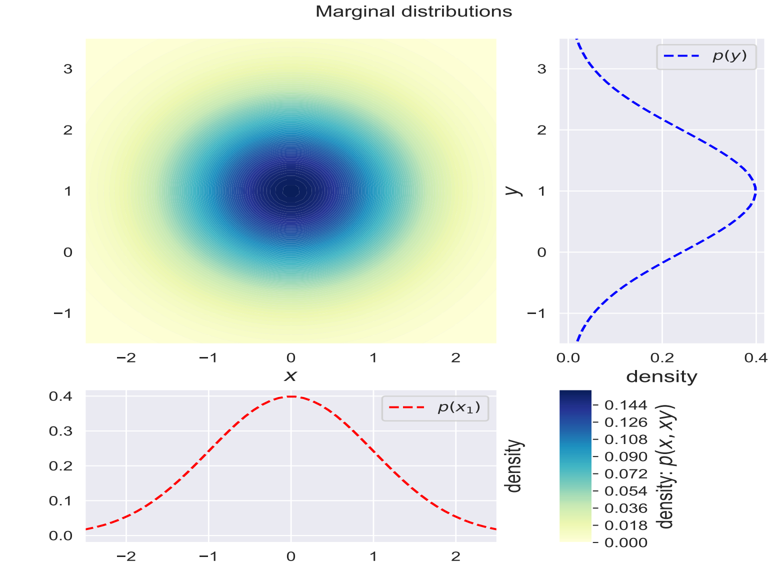
Marginal distributions



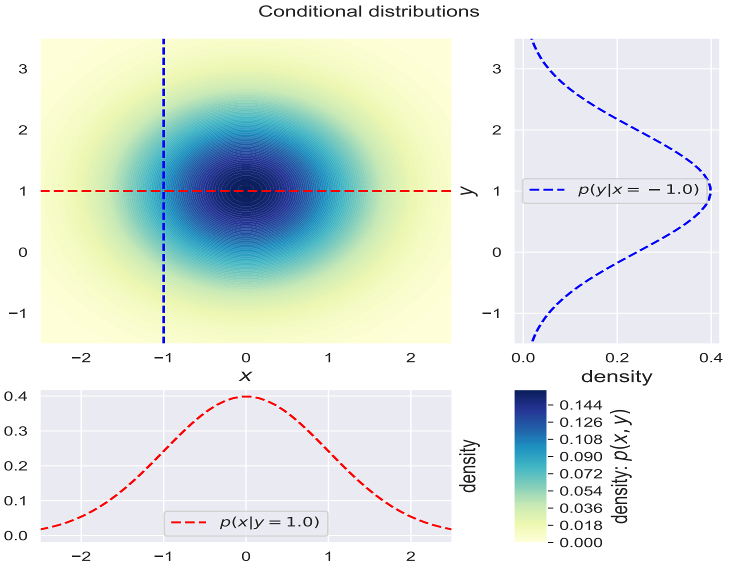
Conditional distributions



Marginal distributions (independent rvs)



Conditional distributions (independent rvs)



James Cussens
james.cussens@bristol.ac.uk