Multivariate Probability

Mathematics for Machine Learning

Lecturer: Matthew Wicker

Material Covered

Models: Linear models, basis expansion, logistic regression, neural networks

Techniques: Least squares estimation, forward AD, reverse AD, computational graphs, gradient descent, convergence, convexity, Lipschitz continuity, Maximum likelihood, maximum a posteriori, LOTUS, change of variables

Settings: Regression, Classification, Density Estimation

This lecture: Multivariate probability, marginalization, basic probability rules, covariance

Last Lecture

Random Variables: Described by probability density/cumulative density functions (or mass functions)

Parameters: Constants that shift the density of the random variable

Expectations: The center of mass for the probability distribution

Variance: The spread of the probability distribution

LOTUS/Change of Var.: Some techniques to manipulate random variables

This Lecture

- Multivariate probability/Joint probability
- Marginal distributions/Marginalization/Product Rule
- Conditional probability
- Manipulating Multivariate distributions

Random Variable Review

$$X \begin{cases} \Omega \text{ - Sample space} \\ \sigma(\Omega) \text{ - Event space} \\ P \text{ - Probability Measure} \end{cases}$$

$P(X \leq x)$

$F(x)$ - Cumulative probability distribution

$f(x)$ - Probability density function

$P(X = x)$

$\mathbb{E}_p[X]$ - Expectation

$\mathrm{Var}_p[X]$ - Variance

Joint random variables

$$p(x, y) = p(\boxed{X} = x, \boxed{Y} = y)$$

A joint random variable (above) can be seen as a multi-dimensional extension of random variables

Joint random variables

$$p(x, y) = p(\boxed{X} = x, \boxed{Y} = y)$$

A joint random variable (above) can be seen as a multi-dimensional extension of random variables. We can have arbitrarily many joint random variables:

$$p(x_1, x_2, \ldots, x_n) = p(\boxed{X_1} = x_1, \boxed{X_2} = x_2, \ldots, \boxed{X_n} = x_n)$$

Joint random variables

$$p(x, y) = p(\boxed{X} = x, \boxed{Y} = y)$$

A joint random variable (above) can be seen as a multi-dimensional extension of random variables. We can have arbitrarily many joint random variables:

$$p(x_1, x_2, \ldots, x_n) = p(\boxed{X_1} = x_1, \boxed{X_2} = x_2, \ldots, \boxed{X_n} = x_n)$$

Joint random variables

We use joint random variables to characterize two or more related probabilistic events.

$$p(x, y) = p(X = x, Y = y)$$

We have seen that we modelled our datasets with an expression very similar to the below:

$$p(x_1, x_2, \ldots, x_n) = p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

But with the i.i.d. assumption we assumed that each r.v. was independent and from the same distribution. So in the context of density estimation it was inappropriate to model a joint

Joint random variables

We use joint random variables to characterize two or more related probabilistic events.

$$p(\boxed{X} = x, \boxed{Y} = y)$$

Your cell-phone signal

The amount of times you repeat yourself

Joint random variables

We use joint random variables to characterize two or more related probabilistic events.

$$p(\boxed{X} = x, \boxed{Y} = y)$$

Your cell-phone signal

The amount of times you repeat yourself

"On a call to my friend I had like 3 or 4 bars but was repeating myself ten to twelve times!"

Joint random variables

"On a call to my friend I had like 3 or 4 bars but was repeating myself ten to twelve times!"

$$p(\boxed{X} = x, \boxed{Y} = y)$$

Your cell-phone signal

The amount of times you repeat yourself

$$p(3 \leq X \leq 4)$$

What function do we use to reason about this?

Joint random variables

"On a call to my friend I had like 3 or 4 bars but was repeating myself ten to twelve times!"

$$p(\boxed{X} = x, \boxed{Y} = y)$$

Your cell-phone signal

The amount of times you repeat yourself

$$p(3 \leq X \leq 4)$$

Cumulative probability density!

Joint random variables

"On a call to my friend I had like 3 or 4 bars but was repeating myself ten to twelve times!"

$$p(\boxed{X} = x, \boxed{Y} = y)$$

Your cell-phone signal

The amount of times you repeat yourself

Joint cumulative density function:

$$F(x_1, \ldots, x_n) := P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

Joint random variables

"On a call to my friend I had like 3 or 4 bars but was repeating myself ten to twelve times!"

$$p(\boxed{X} = x, \boxed{Y} = y)$$

Your cell-phone signal

The amount of times you repeat yourself

A natural way to think about this is as computing the probability inside of a rectangular region

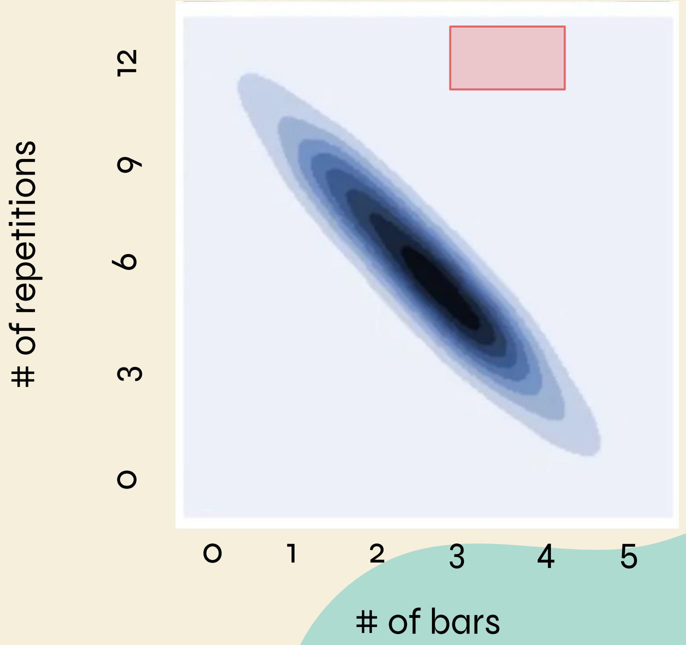
Joint random variables

"On a call to my friend I had like 3 or 4 bars but was repeating myself ten to twelve times!"

$$p(3 \leq X \leq 4, 10 \leq Y \leq 12)$$

In this case, it seems like this occurring has very low probability under our model

$p(x, y)$



Marginal Distribution

We may find ourselves in the case where we have a joint probability density function, but want to know about the distribution of just one of them.

$$p(x, y)$$

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

y

x

	Type 1	Type 2
Malignant	4	8
Benign	7	9

If we know that a patient has a disease but not which specific type we may want to know the probability over the disease status

Marginal Distribution

It is clear to see that in this case, we simply sum along the columns of our dataset and then we can write the probability in the "margin" (hence the name).

$$p(x, y)$$

	Type 1	Type 2	Total	
Malignant	4	8	12	0.42
Benign	7	9	16	0.58

We call the process of summing over the variable we do not care about "marginalization"

Marginal Distribution

$p(x, y)$	Type 1	Type 2	Total		
Malignant	4	8	12	0.42	$= p(X = \text{malig.})$
Benign	7	9	16	0.58	$= p(X = \text{benign})$

We call the process of summing over the variable we do not care about "marginalization"

sum across other variable

$$P(X = x) = \sum_y P(x, y)$$

Marginal Distribution

$P(X = x)$

$p(x, y)$	Type 1	Type 2	Total	
Malignant	4	8	12	0.42
Benign	7	9	16	0.58

We call the process of summing over the variable we do not care about "marginalization"

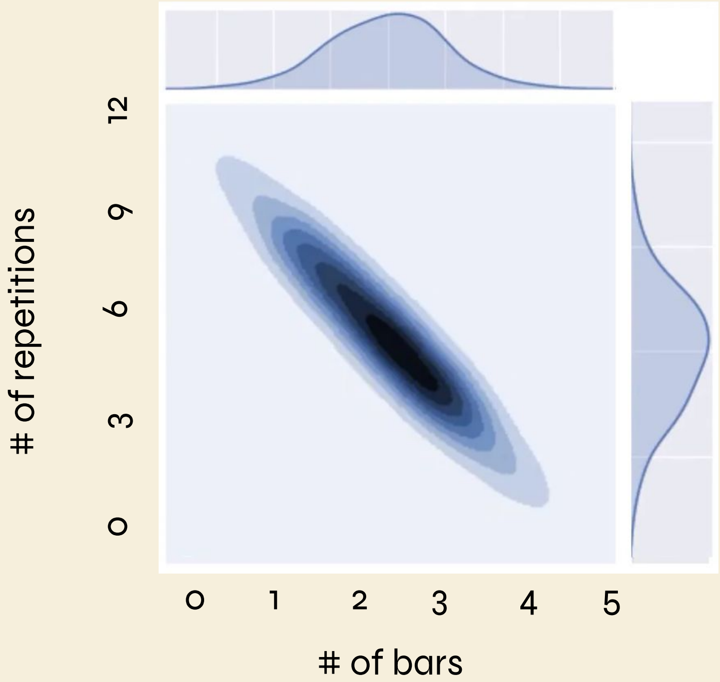
$$P(X = x) = \sum_y P(x, y)$$

$$= \sum_y P(X = x | Y = y) P(Y = y)$$

We will prove this in one second

Marginal Distribution

$$P(X = x) = \sum_{y} P(X = x | Y = y) P(Y = y)$$



Continuous case:

$$\int_{y \in \Omega_y} P(X = x | Y = y) P(Y = y) dy$$

After conducting our study on the number of repetitions in our conversation and the our cellular service, we can use marginalization to get the distribution of bars we have (an average of 2.5)

Marginal Distribution

I have just given this form of the marginal distribution, but let us see a few components that prove this

$$P(X = x) = \sum_y P(X = x|Y = y)P(Y = y)$$

Continuous case:

$$\int_{y \in \Omega_y} P(X = x|Y = y)P(Y = y)dy$$

Conditional Probability

$$P(X \in A | Y \in B) := \frac{\boxed{P(X \in A, Y \in B)}}{P(Y \in B)}$$

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

The conditional probability is defined as the probability distribution of a random variable X given some information information about a jointly distributed random variable.

$P(X,Y) = P(X|Y)\,P(Y)$

Conditional vs. Marginal

$P(X = x)$

$p(x, y)$	Type 1	Type 2	Total	
Malignant	4	8	12	0.42
Benign	7	9	16	0.58
	11	17		
	0.36			
	0.64			

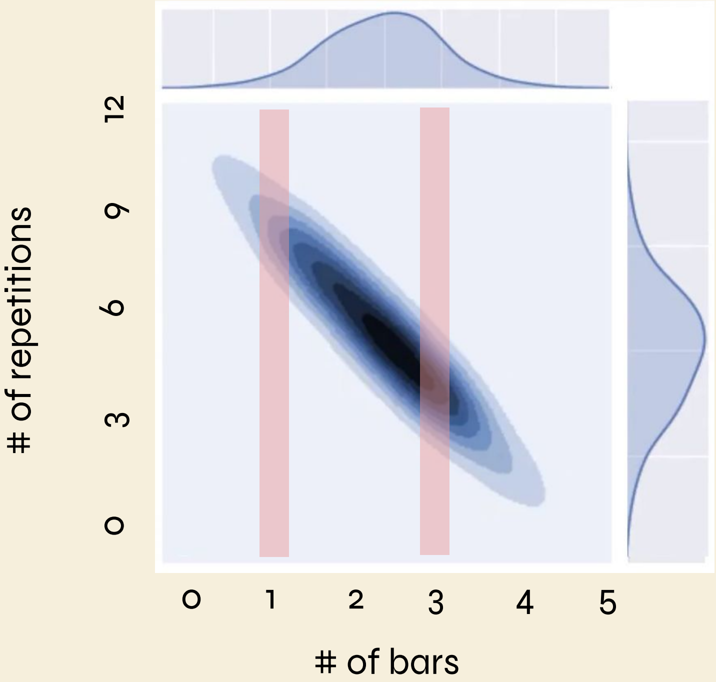
← P(Malig) - not looking at given T1/T2

Marginal distribution

Conditional distribution - only looking at Type 1 -> P(Type 1 | Malig) & P(T1 | Benign)

Conditional Probability

$$P(X \in A | Y \in B) := \frac{P(X \in A, Y \in B)}{P(Y \in B)}$$

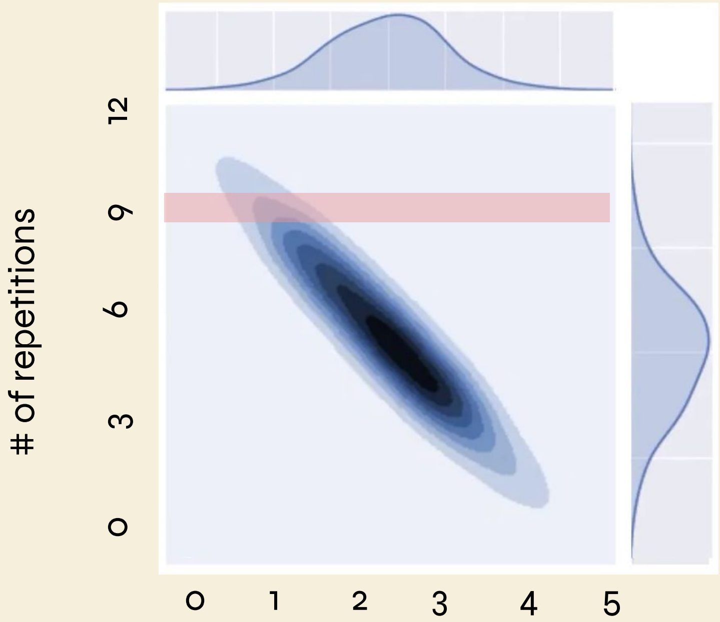


given # bars = 1 or 3

Product rule

Joint distribution = conditional x marginal

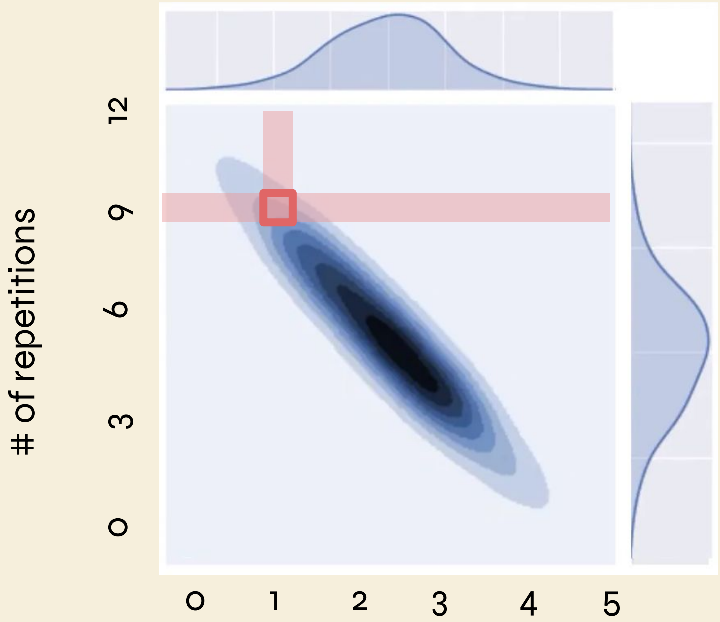
$$P(X \in A, Y \in B) = P(X \in A|y \in B)P(Y \in B)$$



Product rule

Joint distribution = conditional x marginal

$$P(X \in A, Y \in B) = P(X \in A | y \in B) P(Y \in B)$$



Product rule

Joint distribution = conditional x marginal

$$P(X \in A, Y \in B) = P(X \in A | y \in B) P(Y \in B)$$

$$\begin{aligned} P(X = x) &= \sum_y P(x, y) \\ &= \sum_y P(X = x | Y = y) P(Y = y) \end{aligned}$$

Independence

independent

$$X_1 \perp\!\!\!\perp X_2 \iff p(X_1, X_2) = p(X_1)p(X_2)$$

We say that two random variables are independent if and only if their joint distribution can be written as the product of the two densities.

Conditional Independence

$$X_1 \perp\!\!\!\perp X_2 \iff p(X_1, X_2) = p(X_1)p(X_2)$$

We say that two random variables are independent if and only if their joint distribution can be written as the product of the two densities.

$$X_1 \perp\!\!\!\perp X_2 | X_3 \iff p(X_1, X_2 | X_3) = p(X_1 | X_3)p(X_2 | X_3)$$

We say that X_1 and X_2 are conditionally independent if the conditional distributions of X_1 and X_2 given X_3 satisfy the above independence property

Conditional Independence Example

Imagine we have two coins one fair coin and one coin that has heads on both sides. We draw a coin at random and flip it twice. Consider the following events:

A - First coin toss is a heads
B - Second coin toss is a heads
C - We are flipping the fair coin

$P(A, B) \neq P(A)P(B)$ in general. But $P(A, B \mid C) = P(A \mid C)P(B \mid C)$

Multivariate distribution properties

$$X = [X_1, \ldots, X_n]^\top$$

$$\mathbb{E}[X] = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n]]^\top$$

Covariance → $$\mathbb{V}[X] = \Sigma$$

Covariance : matrix, symmetric, PSD

$$\Sigma_{i,j} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$$

Multivariate distribution properties

$$X = [X_1, \ldots, X_n]^\top \qquad \mathbb{E}[X] = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n]]^\top$$

$$\begin{aligned}
\mathbb{E}(X) &= \int_{\mathbb{R}^n} X f_X(X) dX \\
&= \int_{\mathbb{R}^n} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} f_X\left(x_1, \ldots, x_n\right) dx_1 \ldots dx_n
\end{aligned}$$

Multivariate distribution properties

$$X = [X_1, \ldots, X_n]^\top \qquad \mathbb{E}[X] = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n]]^\top$$

$$= \begin{pmatrix} \int_{\mathbb{R}^n} x_1 f_X(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \\ \vdots \\ \int_{\mathbb{R}^n} x_n f_X(x_1, \ldots, x_n)\, dx_1 \ldots dx_n \end{pmatrix}$$

Multivariate distribution properties

$$X = [X_1, \ldots, X_n]^\top \qquad \mathbb{E}[X] = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n]]^\top$$

other vars or constants

$$= \begin{pmatrix} \int_{\mathbb{R}} x_1 f_{X_1}(x_1)\, dx_1 \\ \vdots \\ \int_{\mathbb{R}} x_n f_{X_n}(x_n)\, dx_n \end{pmatrix}$$

Multivariate distribution properties

$$X = [X_1, \ldots, X_n]^\top \qquad \mathbb{E}[X] = [\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n]]^\top$$

$$= \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix}$$

Manipulating rvs: Linearity of expectation

We have established that the expectation of a vector valued random variable is the vector with each element being equal to the expectation of the element. What can we say about these values:

$$\mathbb{E}[aX] \qquad \mathbb{E}[X+Y]$$

$= a\,E(X)$ $\qquad$ $= E[X] + E(Y)$

Manipulating rvs: Linearity of expectation

We have established that the expectation of a vector valued random variable is the vector with each element being equal to the expectation of the element. What can we say about these values:

$$\mathbb{E}[aX] \qquad \mathbb{E}[X+Y]$$

Recall that LOTUS lets us reason about functions of this form so that would be a good starting place!

Manipulating rvs: Linearity of expectation

Using LOTUS and the same line of logic we used to show the expectation of a vector valued rv is the vector of the element-wise expectations we can prove:

$$\mathbb{E}[aX] = a\mathbb{E}[X]$$

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

Conditional Expectations

$$\mathbb{E}[X|Y=y]$$

$$\mathbb{E}[X|Y=y] = \int x \; p(X=x|Y=y)dx$$

Conditional Expectations

$$\mathbb{E}[X|Y = y]$$

$$\mathbb{E}[X|Y = y] = \int x \; p(X = x|Y = y)dx$$

Law of total expectation:

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$$

Law of Total Expectation

$$\mathbb{E}[\mathbb{E}[X|Y]] =$$

$$\int \left(\int xP(X = x|Y = y)\,dx \right) P(Y = y)\,dy$$

By definition

Law of Total Expectation

$$\mathbb{E}[\mathbb{E}[X|Y]] =$$

$$\int \left(\int xP(X = x|Y = y) dx \right) p(Y = y)dy$$

By definition

$$\int \int xP(X = x, Y = y)dxdy$$

Joint = Marg x Cond

Law of Total Expectation

$$\mathbb{E}[\mathbb{E}[X|Y]] =$$

$$\int \left(\int xP(X = x|Y = y) \right) p(Y = y)dy$$

By definition

$$\int \int xP(X = x, Y = y)dxdy$$

Joint = Marg x Cond

$$\int x \left(\underbrace{\int P(X = x, Y = y)dy}_{\text{Marginal for X}} \right) dx$$

Algebra

Law of Total Expectation

$$\mathbb{E}[\mathbb{E}[X|Y]] =$$

$$\int x\left(\int P(X=x, Y=y)dy\right)dx$$

Algebra

$$\int xP(X=x)dx$$

Marginal distribution

Law of Total Expectation

$$\mathbb{E}[\mathbb{E}[X|Y]] =$$

$$\int x\left(\int P(X = x, Y = y)dy\right)dx$$

Algebra

$$\int xP(X = x)dx$$

Marginal distribution

$$\mathbb{E}[X]$$

Definition

4.4 Law of Total Variance

We do not prove this here as it is a bit longer than the proof of the above law of total expectation, but a similar rule holds for variance which we call the law of total variance. This states that:

$$\mathbb{V}[Y] = \mathbb{E}_Y[\mathbb{V}[Y|X]] + \mathbb{V}[\mathbb{E}[Y|X]]$$

Multivariate Gaussian

One of the distributions we will work with most often is the mv Gaussian distribution:

$$f(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2} \underbrace{|\Sigma|^{1/2}}_{\text{sqrt of determinant of covariance matrix}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

Though we have derived the negative log-likelihood of this distribution in previous lectures it can be good practice to do this yourself in your reviews without looking at the notes

Identifying MV Probability in ML

$p(y\|x, \theta)$	Predictive distribution
$\mathbb{E}_{p(y\|x,\theta)}[Y]$	Posterior predictive mean
$\mathbb{V}_{p(y\|x,\theta)}[Y]$	Posterior predictive variance

Identifying MV Probability in ML

$$p(\theta|X, Y)$$

Posterior distribution over model parameters

Next lecture: Bayes Theorem & Math of Beliefs