Machine Learning I Lecture II: Tutorial on probability theory

AMIT DHOMNE

Plan for today

Discrete probability distributions

Multivariate distributions: Joints, conditionals and marginals

Continuous probability distributions

To specify a discrete random variable, we need a sample space and a probability mass function.

- Sample space Ω: Possible 'states' x of the random variable X (outcomes of the experiment, output of the system, measurement). Examples: [on board]
- Discrete random variables either have a finite or countable number of states.
- Events: Possible combinations of states ('subsets of Ω')
- Probability mass function $P(X = x)$: A function which tells us how likely each possible outcome is.

$$P(X = x) = P_X(x) = P(x) \tag{1}$$

$$P(x) \geq 0 \text{ for each } x \tag{2}$$

$$\sum_{x \in \Omega} P(x) = 1 \tag{3}$$

$$P(A) = P(x \in A) = \sum_{x \in A} P(X = x) \tag{4}$$

$$\tag{5}$$

- We write: $X|q \sim \text{Binomial}(q)$
- Bernoulli, Binomial, Multinonomial, Poisson: [on board]

Conditional probability: Updating probabilities after we obtain information.

- Conditional probability: 'Recalculated probability of event A after someone tells you that event A happened.'

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{6}$$

$$P(A \cap B) = P(A|B)P(B) \tag{7}$$

- Examples: Rolls of a die [on board]
- Bayes Rule:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{8}$$

Expectation and variance characterize the mean value of a random variable and its dispersion.

- Expectation (or mean): $E(X) = \sum_x P(X = x)x$
- Expectation of a function: $E(f(X)) = \sum_x P(X = x)f(x)$
- Moments= expectation of power of X: $M_k = E(X^k)$
- Variance: Average (squared) fluctuation from the mean

$$\begin{aligned} \mathrm{Var}(X) &= E((X - E(X))^2) && (9)\\ &= E(X^2) - E(X)^2 && (10)\\ &= M_2 - M_1^2 && (11) \end{aligned}$$

- Standard devation: Square root of variance.
- Illustration and examples: [on board] Aside: Difference between expectation/variance of random variable and empirical average/variance.

Bivariate distributions characterize systems with two observables.

- Example [on board]
- Joint distribution: $P(X = x, Y = y)$, a list of all probabilities of all possible pairs of observations
- Marginal distribution: $P(X = x) = \sum_y P(X = x, Y = y)$
- Conditional distribution: $P(X = x|Y = y) = \frac{P(X=x,Y=y)}{P(y=y)}$
- $X|Y$ has distribution $P(X|Y)$, where $PX(|Y)$ specifies a 'lookup-table' of all possible $P(X = x|Y = y)$

Conditioning and marginalization come up in Bayesian inference ALL the time: 'Condition on what you observe. Marginalize out the uncertainty'.

The importance of conditional probabilities: Interpreting medical tests

	Positive Test	Negative Test	
HIV	475	25	500
no HIV	4975	94525	99500
	5450	94550	100000

[on board]

Source: Statistical Methods for the Social Sciences, Agresti and Finaly, Prentice Hall– not actual data

Expectation and covariance of multivariate distributions:

- Conditional distributions are just distributions which have a (conditional) mean or variance.
- Note: $E(X|Y) = f(Y)$. 'If I tell you what Y is, what is the average value of X?.
- Covariance is the expected value of the product of fluctuations:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= E\left((X - E(X))(Y - E(Y))\right) && (12)\\ &= E(XY) - E(X)E(Y) && (13)\\ \mathrm{Var(X)} &= \mathrm{Cov}(X,X) && (14) \end{aligned}$$

Aside: One common way to construct bivariate random variables is to have a random variable whos parameter is another random variable.

Independence of random variables

- Intuitively, two events are independent if knowing that the first took places tells us nothing about the probability of the second: $P(A|B) = P(A)$
- $P(A)P(B) = P(A \cap B)$
- Two random variables are independent if the joint p.m.f. is the product of the marginals: $P(X = x, Y = y) = P(X = x)P(Y = y)$.
- If X and Y are independent, we write $X \perp Y$. Knowing the value of X does not tell us anything about Y.
- If X and Y are independent, $\text{Cov}(X, Y) = 0$.

Aside: Mutual information is a measure of how 'non-independent' two random variables are.

Multivariate distributions are the same as bivariate distributions, just with more dimensions.

- $\mathbf{X}, \mathbf{x}$ are vector valued.
- Mean: $E(\mathbf{X}) = \sum_{\mathbf{x}} \mathbf{x} P(\mathbf{x})$
- Covariance matrix:

$$\mathrm{Cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) \tag{15}$$

$$\mathrm{Cov}(\mathbf{X}) = E(\mathbf{X}\mathbf{X}^\top) - E(\mathbf{X})E(\mathbf{X})^\top \tag{16}$$

- Conditional and marginal distributions: Can define and calculate any (multi or single-dimensional) marginals or conditional distributions we need: $P(X_1)$, $P(X_1, X_2)$, $P(X_1, X_2, X_3|X_4)$, etc..

Continous random variables

- A random variable X is continuous if its sample space X is uncountable.
- In this case, $P(X = x) = 0$ for each x.
- If $p_X(x)$ is a probability density function for X, then

$$P(a < X < b) = \int_a^b p(x)dx \tag{17}$$

$$P(a < X < a + dx) \approx p(a) \cdot dx \tag{18}$$

- The cumulative distribution function is $F_X(x) = P(X < x)$. We have that $p_X(x) = F'(x)$, and $F(x) = \int_{-\infty}^{x} p(s)ds$.
- More generally: If A is an event, then

$$P(A) = P(X \in A) = \int_{x \in A} p(x)dx \tag{19}$$

$$P(\Omega) = P(X \in \Omega) = \int_{x \in \Omega} p(x)dx = 1 \tag{20}$$

- Example: Uniform, Exponential, Beta [on board]

People will often say probability when they mean probability density.

- Probability density functions do not satisfy the definitions of probability (e.g. they can bigger than 1). However, people (including your lecturer) will often be sloppy and write things like $P(X = x)$ and say 'the probability of X' when they really mean 'the probability density of X evaluated at x'.
- Similarly, people (including your lecturer) will often be sloppy and write integrals or say 'we need to integrate our X' when they write down general formulas– if these formulas are applied to discrete random variables, the integrals would need to be replaced by sums.
- This might be bad practice, but it is usually clear from the context whether a random variable is discrete or continuous. In addition, it is good preparation for reading papers—many machine learning papers are very sloppy about usage of these terms.

Mean, variance, and conditioning on events are the same as the discrete case, just with sums replaced by integrals.

- Mean: $E(X) = \int_x x \cdot p(x)dx$
- Variance: $\mathrm{Var}(X) = E(X^2) - E(X)^2$
- Example: Uniform, Exponential [on board]
- If X has pdf $p(x)$, then $X|(X \in A)$ has pdf

$$p_{X|A}(x) = \frac{p(x)}{P(A)} = \frac{p(x)}{\int_{x \in A} p(x)dx} \tag{21}$$

- Only makes sense if $P(A) > 0$!
- Example: Uniform, Exponential [on board]

Bivariate continuous distributions: Marginalization, Conditioning and Independence

- $p_{X,Y}(x, y)$, joint probablity density function of X and Y
- $\int_x \int_y p(x, y)dxdy = 1$
- Marginal distribution: $p(x) = \int_{-\infty}^{\infty} p(x, y)dy$
- Conditional distribution: $p(x|y) = \frac{p(x,y)}{p(y)}$
- Note: $P(Y = y) = 0$! Formally, conditional probability in the continuous case can be derived using infinitesimal events.
- Independence: X and Y are independent if $p_{X,Y}(x, y) = p_X(x)p_Y(y)$