Machine Learning and Intelligent Systems

A Quick Review on Probability

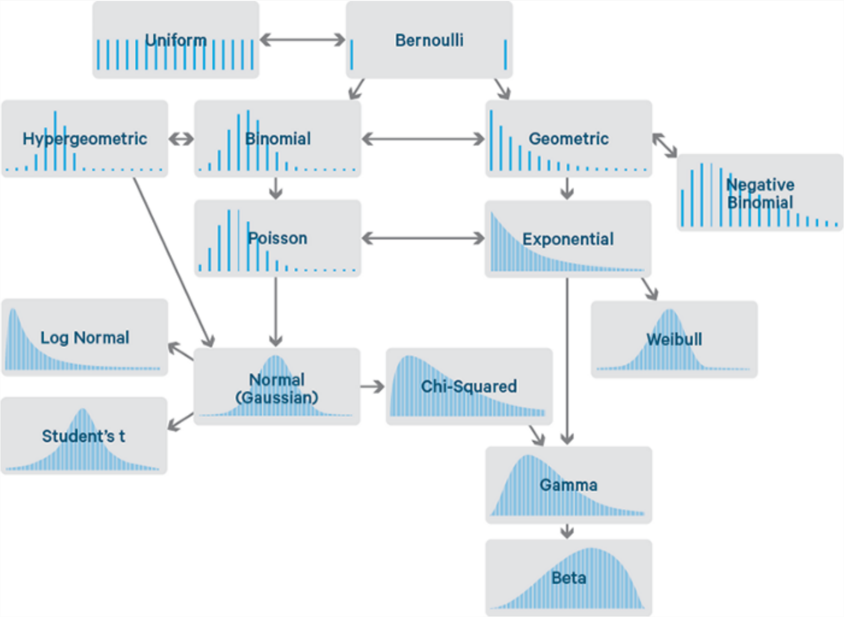
Maria A. Zuluaga

Pre-lecture material

EURECOM - Data Science Department

Table of contents

Distributions



Source: `https://bit.ly/2EQv2cf`

Probability Density Function and Probability Mass Function

- The **probability density function (PDF)**, or density of a continuous random variable, is a function that describes the relative **likelihood** for this random variable to take on a given value.
- It is the primary means of defining a probability distribution.

Probability Density Function and Probability Mass Function

- The **probability density function (PDF)**, or density of a continuous random variable, is a function that describes the relative **likelihood** for this random variable to take on a given value.
- It is the primary means of defining a probability distribution.

- A **probability mass function (PMF)** is a function that gives the probability that a discrete random variable is exactly equal to some value.
- It is often the primary means of defining a discrete probability distribution.

Some Common Distributions

Gaussian or Normal Distribution

- The Gaussian is the most widely used distribution for continuous variables.
- For univariate variables z it is governed by two parameters: the mean μ and the variance $\sigma^2 > 0$.

Gaussian or Normal Distribution

- The Gaussian is the most widely used distribution for continuous variables.
- For univariate variables z it is governed by two parameters: the mean μ and the variance $\sigma^2 > 0$.
- For D-dimensional vectors, it is governed by $\boldsymbol{\mu}$ and a $D \times D$ covariance matrix Σ (symmetric and positive-definite).

Gaussian or Normal Distribution

- The Gaussian is the most widely used distribution for continuous variables.
- For univariate variables z it is governed by two parameters: the mean μ and the variance $\sigma^2 > 0$.
- For D-dimensional vectors, it is governed by $\boldsymbol{\mu}$ and a $D \times D$ covariance matrix Σ (symmetric and positive-definite).

Univariate:

$$N(z|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(z-\mu)^2}{2\sigma^2} \right\}$$

$$\mathbb{E}[z] = \mu$$

$$var[z] = \sigma^2$$

Gaussian or Normal Distribution

- The Gaussian is the most widely used distribution for continuous variables.
- For univariate variables z it is governed by two parameters: the mean μ and the variance $\sigma^2 > 0$.
- For D-dimensional vectors, it is governed by $\boldsymbol{\mu}$ and a $D \times D$ covariance matrix Σ (symmetric and positive-definite).

Multivariate:

$$\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{z}-\boldsymbol{\mu})\right\}$$

$$\mathbb{E}[\boldsymbol{z}] = \boldsymbol{\mu}$$

$$cov[\boldsymbol{z}] = \Sigma$$

Bernoulli Distribution

- Distribution for a single binary variable $z \in \{0, 1\}$

Bernoulli Distribution

- Distribution for a single binary variable $z \in \{0, 1\}$
- Used for representing a coin toss

Bernoulli Distribution

- Distribution for a single binary variable $z \in \{0, 1\}$
- Used for representing a coin toss
- It is governed by a single continuous parameter $p \in [0, 1]$ that represents the probability of $z = 1$

Bernoulli Distribution

- Distribution for a single binary variable $z \in \{0, 1\}$
- Used for representing a coin toss
- It is governed by a single continuous parameter $p \in [0, 1]$ that represents the probability of $z = 1$

$$
\begin{aligned}
Bern(z|p) &= p^z (1-p)^{1-z} \\
\mathbb{E}[z] &= p \\
var[z] &= p(1-p)
\end{aligned}
$$

Binomial Distribution

- Gives the probability of observing m occurrences of $z = 1$ in a set of N samples from a Bernoulli distribution, where the probability of observing $z = 1$ is $z \in \{0, 1\}$

Binomial Distribution

- Gives the probability of observing m occurrences of $z = 1$ in a set of N samples from a Bernoulli distribution, where the probability of observing $z = 1$ is $z \in \{0, 1\}$
- Last lecture example with 10 coin tosses follows the binomial distribution. One toss follows Bernoulli

Binomial Distribution

- Gives the probability of observing m occurrences of $z = 1$ in a set of N samples from a Bernoulli distribution, where the probability of observing $z = 1$ is $z \in \{0, 1\}$
- Last lecture example with 10 coin tosses follows the binomial distribution. One toss follows Bernoulli

$$Bin(m|N, p) = \begin{pmatrix} N \\ m \end{pmatrix} p^m (1-p)^{1-m}$$

$$\mathbb{E}[m] = Np$$

$$var[m] = Np(1-p)$$

- The multinomial distribution is a generalization of the binomial distribution

Multinomial Distribution

- Multivariate generalization of the binomial that gives the distribution over counts m_k for a K-state discrete variable $\boldsymbol{z}$ to be in state k given a total number of observations N.

$$\boldsymbol{z} = [z_1, z_2, z_3, \ldots, z_k], \quad z_k \in \{0, 1\}$$

Multinomial Distribution

- Multivariate generalization of the binomial that gives the distribution over counts m_k for a K-state discrete variable $\boldsymbol{z}$ to be in state k given a total number of observations N.

$$\boldsymbol{z} = [z_1, z_2, z_3, \dots, z_k], \quad z_k \in \{0, 1\}$$

- When $K = 2$ and $N = 1$, it is Bernoulli distribution. When $K = 2$ and $N > 1$ it is the binomial distribution.

Multinomial Distribution

- Multivariate generalization of the binomial that gives the distribution over counts m_k for a K-state discrete variable $\boldsymbol{z}$ to be in state k given a total number of observations N.

$$\boldsymbol{z} = [z_1, z_2, z_3, \ldots, z_k], \quad z_k \in \{0, 1\}$$

- When $K = 2$ and $N = 1$, it is Bernoulli distribution. When $K = 2$ and $N > 1$ it is the binomial distribution.

$$Mult(m_1, m_2, m_3, \ldots, m_K | \boldsymbol{\mu}, N) = \left(\begin{array}{c} N \\ m_1 m_2 \ldots m_M \end{array} \right) \prod_{k=1}^{M} p_k^{m_k}$$

$$\mathbb{E}[m_k] = Np_k$$

$$var[m_k] = Np_k(1 - p_k)$$

$$cov[m_j m_k] = -Np_j p_k$$

with $\boldsymbol{\mu} = (p_1, \ldots, p_K)^T$

MLE and the Coin Toss Problem (revisited)

Formulation:

- You ask yourself, "What is the probability that this coin comes up heads when I toss it?"
- You toss it $n = 10$ times and obtain the following sequence of outcomes:

$$\mathcal{D} = \{H, T, T, H, H, H, T, T, T, T\}.$$

- Based on these samples, how would you estimate $P(H)$?

MLE and the Coin Toss Problem (revisited)

Formulation:

- You ask yourself, "What is the probability that this coin comes up heads when I toss it?"
- You toss it $n = 10$ times and obtain the following sequence of outcomes:

$$\mathcal{D} = \{H, T, T, H, H, H, T, T, T, T\}.$$

- Based on these samples, how would you estimate $P(H)$?

Let's try to formalize it:

- We want to find an expression for P(H)
- What is P(H)?
- This problem can be solved, through MLE, in two different ways

Approach 1: Bernoulli trials

- Lets define $z \in \{0, 1\}$ a random binary variable representing the outcome of **a single coin toss**.
- Heads: $z = 1$, Tails: $z = 0$

Approach 1: Bernoulli trials

- Lets define $z \in \{0, 1\}$ a random binary variable representing the outcome of **a single coin toss**.
- Heads: $z = 1$, Tails: $z = 0$
- Thus we can formalize $\mathcal{D} = \{H, T, T, H, H, H, T, T, T, T\}$, as the set $\mathcal{D} = \{z_1, \ldots, z_N\}$ of observed values of z.

Approach 1: Bernoulli trials

- Lets define $z \in \{0, 1\}$ a random binary variable representing the outcome of **a single coin toss**.
- Heads: $z = 1$, Tails: $z = 0$
- Thus we can formalize $\mathcal{D} = \{H, T, T, H, H, H, T, T, T, T\}$, as the set $\mathcal{D} = \{z_1, \ldots, z_N\}$ of observed values of z.

- We just saw that the outcome of a **single coin toss** follow the Bernoulli distribution
- Parameter:
  - $0 \leq p \leq 1$ - The probability of heads
  - $q = 1 - p$ - the probability of tails (can be obtained through p)

Approach 1: Bernoulli trials

- Lets define $z \in \{0, 1\}$ a random binary variable representing the outcome of **a single coin toss**.
- Heads: $z = 1$, Tails: $z = 0$
- Thus we can formalize $\mathcal{D} = \{H, T, T, H, H, H, T, T, T, T\}$, as the set $\mathcal{D} = \{z_1, \dots, z_N\}$ of observed values of z.

- We just saw that the outcome of a **single coin toss** follow the Bernoulli distribution
- Parameter:
  - $0 \leq p \leq 1$ - The probability of heads
  - $q = 1 - p$ - the probability of tails (can be obtained through p)

Goal: Use MLE to find p

Approach 1: Bernoulli trials

With all these elements we can now construct the likelihood function, under the assumption that the observations (the coin tosses) are independent, so that

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(z_i|\theta)$$

The log-likelihood is given by (see MLE slides):

$$\begin{aligned}
\log p(\mathcal{D}|p) &= \sum_{i=1}^{N} \log p(z_i|\theta) \\
&= \sum_{i=1}^{N} \log \left(p^{z_i}(1-p)^{1-z_i}\right) \\
&= \sum_{i=1}^{N} z_i \log p + (1-z_i)\log(1-p)
\end{aligned}$$

Approach 1: Bernoulli trials

Using the expression found for $\log p(\mathcal{D}|p)$ we can find p_{MLE} by:

$$p_{MLE} = \arg\min_{p} \sum_{i=1}^{N} z_i \log p + (1 - z_i) \log(1 - p)$$

Approach 1: Bernoulli trials

Using the expression found for $\log p(\mathcal{D}|p)$ we can find p_{MLE} by:

$$p_{MLE} = \arg\min_{p} \sum_{i=1}^{N} z_i \log p + (1 - z_i) \log(1 - p)$$

How?

$$\frac{\partial}{\partial p} \left(\sum_{i=1}^{N} z_i \log p + (1 - z_i) \log(1 - p) \right) = 0$$

and solving for p.

(quick) Exercise: Your task to complete it and find p.

Approach 2: Binomial distribution

- We know that the binomial distribution is nothing else than repeated Bernoulli trials
- Parameters:
    - N: The number of trials $\longrightarrow N = n_H + n_T$ in our example
    - p: The probability of heads

Approach 2: Binomial distribution

- We know that the binomial distribution is nothing else than repeated Bernoulli trials
- Parameters:
  - N: The number of trials $\longrightarrow N = n_H + n_T$ in our example
  - p: The probability of heads

- In this case, we express the observations as $\mathcal{D} \sim Bin(N, p)$

Approach 2: Binomial distribution

- We know that the binomial distribution is nothing else than repeated Bernoulli trials
- Parameters:
  - N: The number of trials $\longrightarrow N = n_H + n_T$ in our example
  - p: The probability of heads

- In this case, we express the observations as $\mathcal{D} \sim Bin(N, p)$

Goal: Use MLE to find p

Approach 2: Binomial distribution

- In this case, we express the observations as $\mathcal{D} \sim Bin(N, p)$
- What does this imply?

$$p(\mathcal{D}|\theta) = \begin{pmatrix} n_H + n_T \\ n_H \end{pmatrix} \theta^{n_H} (1-\theta)^{n_T}$$

with $\theta \Rightarrow p$

Approach 2: Binomial distribution

- In this case, we express the observations as $\mathcal{D} \sim Bin(N, p)$
- What does this imply?

$$p(\mathcal{D}|\theta) = \binom{n_H + n_T}{n_H} \theta^{n_H} (1-\theta)^{n_T}$$

  with $\theta \Rightarrow p$

- When we plug-in this expression into our likelihood estimator the product (or sum) disappears
- **Why?**

Approach 2: Binomial distribution

This leads us to:

$$p_{MLE} = \arg\min_{p} \log\left(\binom{n_H + n_T}{n_H} p^{n_H} (1-p)^{n_T}\right)$$

which is solved through:

$$\frac{\partial}{\partial p}\left(n_H \log p + n_T \log(1-p)\right) = 0$$

Approach 2: Binomial distribution

This leads us to:

$$
\begin{aligned}
p_{MLE} &= \arg\min_{p} \log\left(\binom{n_H + n_T}{n_H} p^{n_H} (1-p)^{n_T}\right) \\
&= \arg\min_{p} \log\left(\cancel{\binom{n_H + n_T}{n_H}} p^{n_H} (1-p)^{n_T}\right)
\end{aligned}
$$

which is solved through:

$$
\frac{\partial}{\partial p}\left(n_H \log p + n_T \log(1-p)\right) = 0
$$

Exercise: Show that this leads to the same results as with Approach 1.

Approach 2: Binomial distribution

This leads us to:

$$
\begin{aligned}
p_{MLE} &= \arg\min_p \ \log\left(\binom{n_H + n_T}{n_H} p^{n_H}(1-p)^{n_T}\right) \\
&= \arg\min_p \ \log\left(\cancel{\binom{n_H + n_T}{n_H}} p^{n_H}(1-p)^{n_T}\right) \\
&= \arg\min_p \ n_H \log p + n_T \log(1-p)
\end{aligned}
$$

which is solved through:

$$
\frac{\partial}{\partial p}\left(n_H \log p + n_T \log(1-p)\right) = 0
$$

Exercise: Show that this leads to the same results as with Approach 1.

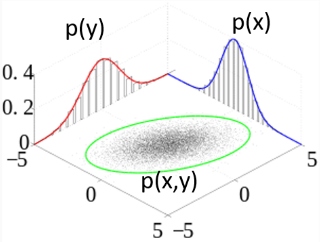
Joint and Conditional Probabilities

Joint Probabilities

- The supervised learning problem (regression) has an input $\mathbf{X}$ and the corresponding target output vector $\mathbf{y}$ with the goal to predict y given a new value $\mathbf{x}$.
- The joint probability distribution $p(\mathbf{x}, y)$ provides a view on the uncertainty of these variables.

Joint Probabilities

- The supervised learning problem (regression) has an input $\mathbf{X}$ and the corresponding target output vector $\mathbf{y}$ with the goal to predict y given a new value $\mathbf{x}$.
- The joint probability distribution $p(\mathbf{x}, y)$ provides a view on the uncertainty of these variables.



Source: `https://en.wikipedia.org/wiki/Joint_probability_distribution`

Joint probability: For two discrete random variables, X and Y, $\mathrm{P}(X = x, Y = y)$ is the probability that random variable X has value x and random Y has value y.

Joint density function: For two continuous random variables, x and y, $p(x, y)$ is the joint density function (pdf).

Conditional Probabilities

- When variables are dependent it is possible to work with conditioning
- Example: Probability of breaking the world marathon record (B=1) given that the temperature will be above 30 (A=1)

$$P(B = 1|A = 1)$$

- Conditional PDF example

$$p(y_i|\mathbf{x}_i; \mathbf{w}, \sigma^2)$$

- Conditional probability example

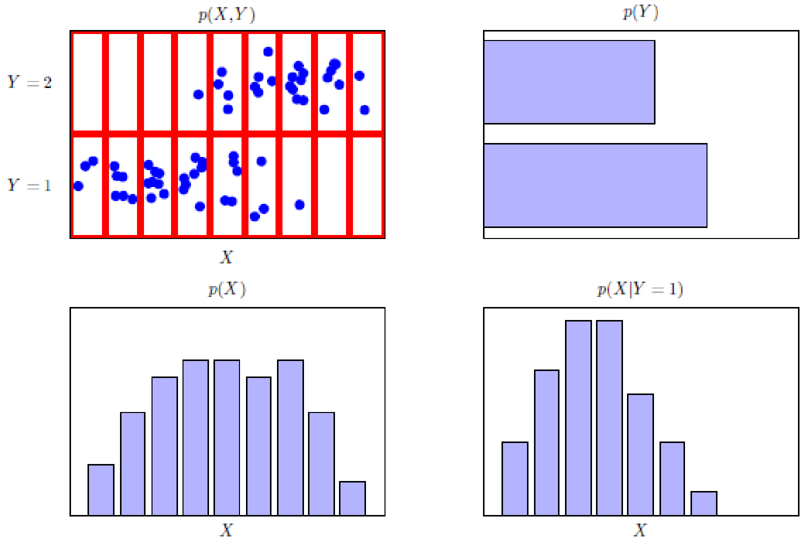
$$P(9 \leq y_i \leq 9.8|\mathbf{x}_i; \mathbf{w}, \sigma^2)$$

The Rules of Probability

	Discrete variables	Continuous variables
Sum rule	$P(X) = \sum_Y P(X, Y)$	$p(x) = \int_y p(x, y)dy$
Product rule	$P(X, Y) = P(Y\|X)P(X)$	$p(x, y) = p(y\|x)p(x)$

- $P(X, Y)$: Joint probability.
- $P(Y|X)$: Conditional probability, e.g. the probability of Y given X.
- $P(X)$: Marginal probability, e.g. the probability of X.

An Illustration



Source: Figure 1.11 from PRML - Bishop

Bayes' Theorem

Using the product rule and the symmetry property $P(X, Y) = P(Y, X)$ we obtain the following relationship among conditional probabilities:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}, \tag{1}$$

which is known as the **Bayes'theorem**.

Bayes' Theorem

Using the product rule and the symmetry property $P(X, Y) = P(Y, X)$ we obtain the following relationship among conditional probabilities:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}, \tag{1}$$

which is known as the **Bayes'theorem**.

Using the sum rule, the denominator in Bayes' theorem can be expressed in terms of the quantities appearing in the numerator:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{\sum_Y P(X|Y)P(Y)} \tag{2}$$

The Elements in the Bayes' theorem

Quantity	Name	Interpretation
$P(Y)$	Prior probability of Y	Probability of a hypothesis Y without any additional prior information
$P(X\|Y)$	Likelihood	Probability of observing the new evidence, given the initial hypothesis
$P(Y\|X)$	Posterior probability	Quantity of interest. Probability of Y given the evidence X
$P(X)$	Evidence or marginal likelihood	Total probability of observing the evidence

Statistical Properties

Expectations & Covariances

Now, let' have a quick recap on some important statistical properties

Expectations & Covariances

Now, let' have a quick recap on some important statistical properties

Expectation: The average of a large number of independent realizations of a random variable:

- <u>Discrete:</u> $\mathbb{E}[X] = \sum_X XP(X)$
- <u>Continuous:</u> $\mathbb{E}[x] = \int_x xp(x)dx$

In either case, if have a finite number N of points drawn from the probability distribution or probability density, then the expectation can be approximated as:

$$\mathbb{E}[x] \simeq \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{(same for discrete case)}$$

Expectations & Covariance

Conditional Expectation: Expected value of y given x

$$\mathbb{E}[y|x] = \int_y y p(y|x) dy \quad \text{(analogous in the discrete case)}$$

Expectations & Covariance

Conditional Expectation: Expected value of y given x

$$\mathbb{E}[y|x] = \int_y yp(y|x)dy \quad \text{(analogous in the discrete case)}$$

Covariance: A measure of joint variability of two variables:

$$cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[Y])(Y - \mathbb{E}[X])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Expectations & Covariance

During the last lecture, we saw how to predict a new point once the model parameters $(\hat{\mathbf{w}}, \sigma^2)$ were estimated:

$$\hat{y}_{new} = \mathbb{E}[y_{new}|\mathbf{x}^*; \mathbf{w}, \sigma^2] = \mathbf{x}^* \hat{\mathbf{w}}$$

Expectations & Covariance

During the last lecture, we saw how to predict a new point once the model parameters $(\hat{\mathbf{w}}, \sigma^2)$ were estimated:

$$\hat{y}_{new} = \mathbb{E}[y_{new} | \mathbf{x}^*; \mathbf{w}, \sigma^2] = \mathbf{x}^* \hat{\mathbf{w}}$$

Exercise: Study the statistical properties of the linear regression estimates $\hat{\mathbf{w}}$, σ: $\mathbb{E}[\hat{\mathbf{w}}]$, $\mathbb{E}[\sigma]$, $cov(\hat{\mathbf{w}})$

References

Further Reading and Useful Material

Source	Notes
Pattern Recognition and Machine Learning	Ch. 2
Bayes' Rule tutorial	Link
Review Notes on Probability	Link