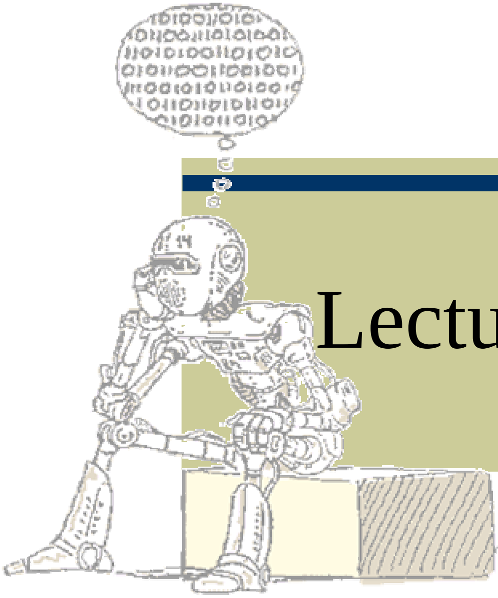


Lecture III: Statistics & probability theory

Overview

- random variables (discrete & continuous)
- distributions (discrete & continuous)
- expected values, moments
- joint distributions, conditional distributions, independence
- Bayes Rule

Note: *Probability theory and distributions form the basis for explanation of data and their generative mechanisms.*

Random Variables

- **A random variable is a random number determined by chance, or more formally, drawn according to a probability distribution**
  - the probability distribution can be given by the physics of an experiment (e.g., throwing dice)
  - the probability distribution can be synthetic
  - discrete & continuous random variables
- **Typical random variables in Machine Learning Problems**
  - the input data
  - the output data
  - noise
- **Important concept in learning: *The data generating model***
  - e.g., what is the data generating model for: i) throwing dice, ii) regression, iii) classification, iv) for visual perception?

Discrete Probability Distributions

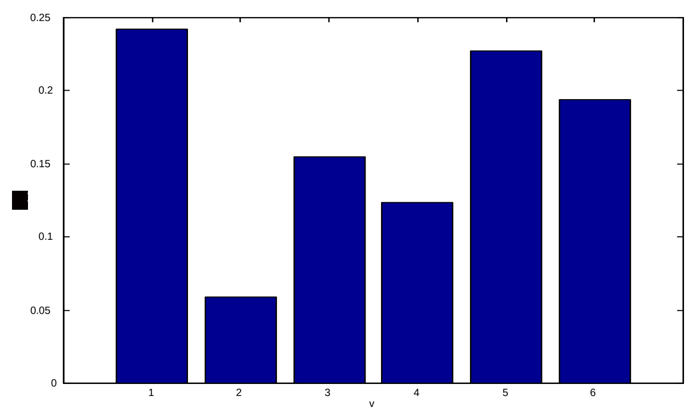
- The random variables only take on **discrete** values
  - e.g., throwing dice: possible values

$$v_i \in \{1, 2, 3, 4, 5, 6\}$$

- The probabilities sum to 1

$$\sum_i P(v_i) = 1$$

- Discrete distributions are particularly important in classification
- Probability Mass Function or Frequency Function (normalized histogram)



A “non fair” die

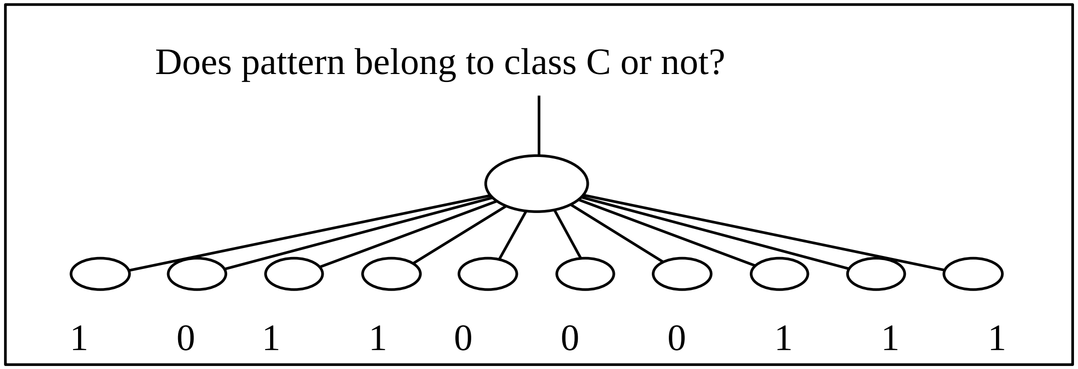
Classic Discrete Distributions (I)

Bernoulli Distribution

- A Bernoulli random variable takes on only two values, i.e., 0 and 1.
- $P(0)=p$ and $P(1)=1-p$, or in compact notation:

$$P(x) = \begin{cases} p^x (1-p)^{1-x}, & \text{if } x = 0 \text{ or } x = 1 \\ 0, & \text{otherwise} \end{cases}$$

- Bernoulli distributions are naturally modeled by sigmoidal activation functions in neural networks (Bishop, Ch.1 & Ch.3) with binary inputs.



Classic Discrete Distributions (II)

Binomial Distribution

- Like Bernoulli distribution: binary input variables: 0 or 1, and probability $P(0)=p$ and $P(1)=1-p$
- What is the probability of k successes, $P(k)$, in a series of n independent trials? (n>=k)
- $P(k)$ is a binomial random variable:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{where} \quad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- Binomial variables are important for density estimation networks, e.g. "what is the probability that k data points fall into region R?" (Bishop, Ch.2)
- Bernoulli distribution is a subset of binomial distribution (i.e., n=1)

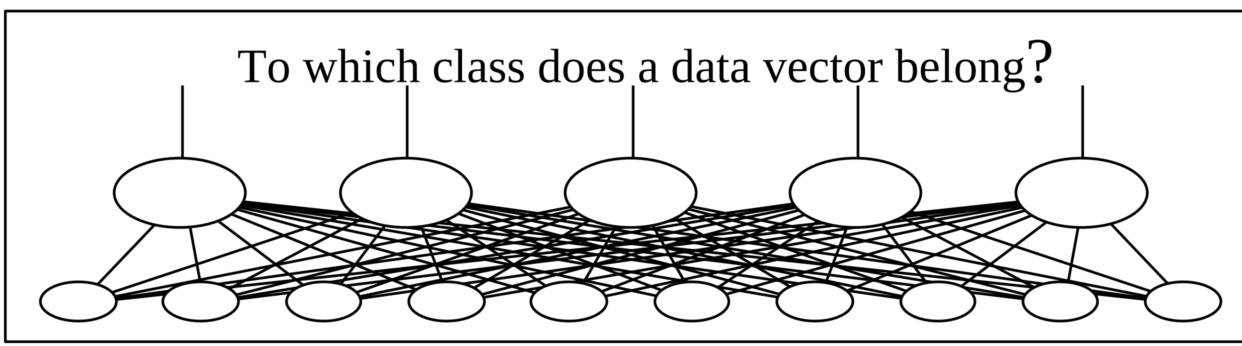
Classic Discrete Distributions (III)

Multinomial Distribution

- A generalization of the binomial distribution to multiple outputs (i.e., multiple classes can be categorized instead of just one class).
- n independent trials can result in one of r types of outcomes, where each outcome c_r has a probability $P(c_r)=p_r$(p_r=1).
- What is the probability $P(n_1,n_2,\ldots,n_r)$, i.e., the probability that in n trials, the frequency of the r classes is $(n_1,n_2,\ldots,n_r)$? This is a multinomial random variable:

$$P(n_1,\ldots,n_r) = \binom{n}{n_1 n_2 \ldots n_r} p_1^{n_1} p_2^{n_2} \ldots p_r^{n_r} \text{ where } \binom{n}{n_1 n_2 \ldots n_r} = \frac{n!}{n_1! n_2! \ldots n_r!}$$

- The multinomial distribution plays an important role in multi-class classification (where n=1).



Classic Discrete Distributions (IV)

Poisson Distribution

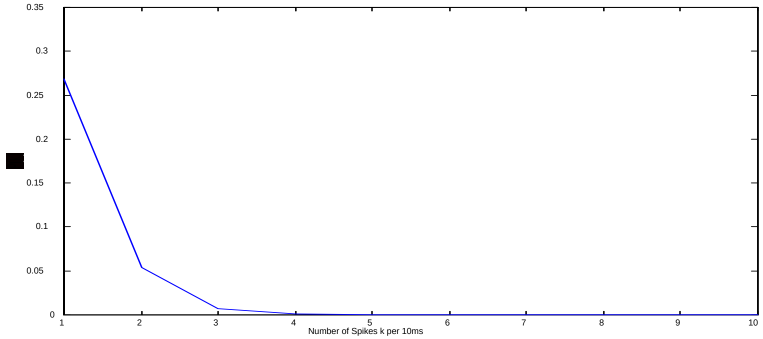
- The Poisson distribution is binomial distribution where the number of trials *n* goes to infinity, and the probability of success on each trial, *p*, goes to zero, such that $np=\lambda$.

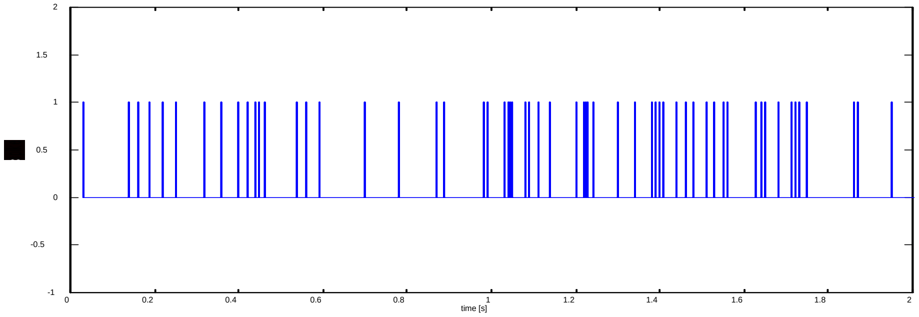
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

- Poisson distributions are an important model for the firing characteristics of biological neurons. They are also used as an approximation to binomial variables with small *p*.

Poisson Distribution (cont'd)

- Example: What is the Poisson distribution of neuronal firing of a cerebellar Purkinje cell in a 10ms interval?
  - we know that the average firing rate of a pyramidal cell is 40Hz
  - λ=40Hz*0.01s=0.4
  - note that approximation only works if probability of spiking is small in the considered interval





Continuous Probability Distributions

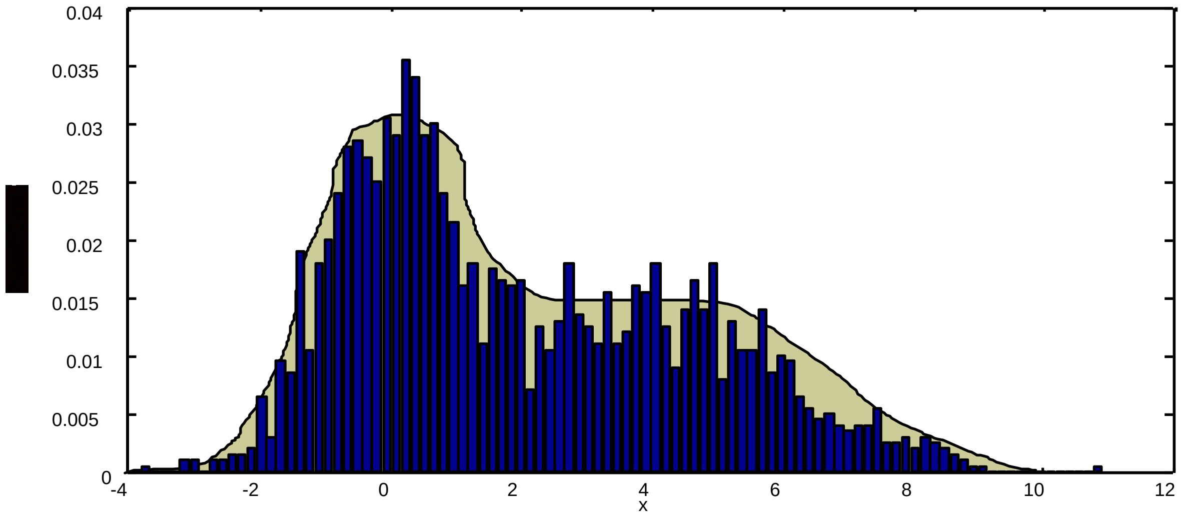
- Random variables take on **real values**.
- Continuous distributions are discrete distributions where the number of discrete values goes to infinity while the probability of each discrete value goes to zero.
- Probabilities become densities.
- Probability density integrates to 1.

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

- Continuous distributions are particularly important in regression.

Continuous Probability Distributions (cont'd)

- Probability Density Function *p*(x)



- Probability of an event:

$$P(a < x < b) = \int_a^b p(x)dx$$

Classic Cont. Distributions (I)

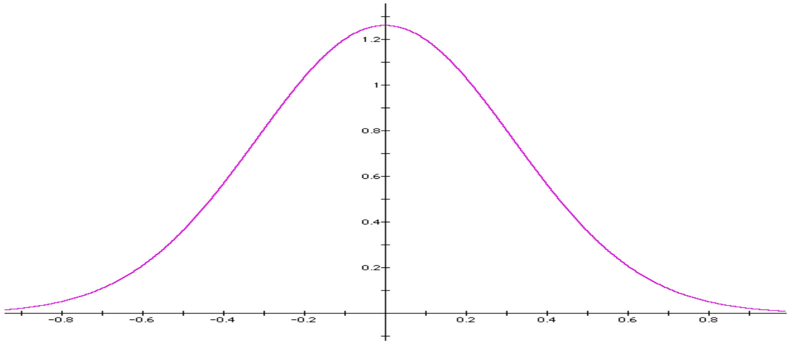
Normal Distribution

- The most important continuous distribution

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu) \right)$$

- Also called Gaussian distribution after C.F.Gauss who proposed it
- Justified by the Central Limit Theorem:
  - roughly: “if a random variable is the sum of a large number of independent random variables, it is approximately normally distributed”
  - Many observed variables are the sum of several random variables
- Shorthand:

$$x \sim N(\mu, \Sigma)$$



Classic Cont. Distribution (II)

The Exponential Family

- A large class of distributions that are all analytically appealing. Why? Because taking the log() of them decomposes them into simple terms.

$$p(\mathbf{x}) = \exp\left(\frac{x\theta - b(\theta)}{a(\phi)} + c(x,\phi)\right)$$

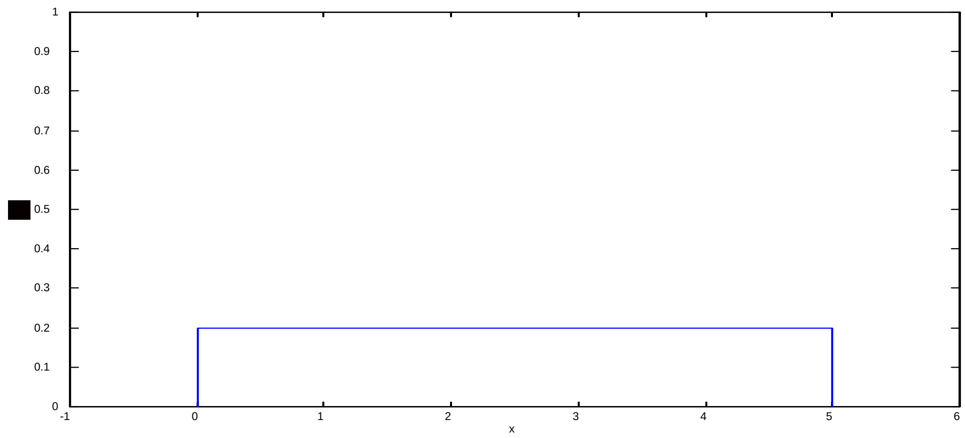
for some specific functions a(), b(), and c(), and parameter vectors θ and ϕ.

- All members are unimodal.
- However, there a many “daily”-life distributions that are not captured by the exponential family.
- **Example distribution in the family**: Univariate Gaussian, Exponential distribution, Rayleigh distribution, Maxwell distribution, Gamma distribution, Beta distribution, Poisson distribution, Binomial distribution, Multinomial distribution.

Classic Cont. Distributions (III)

Uniform Distribution

- All data is equally probable within a bounded region R, $p(x)=1/R$.



Uniform distributions play a very important role in machine learning based on information theory and entropy methods.

Expected Values

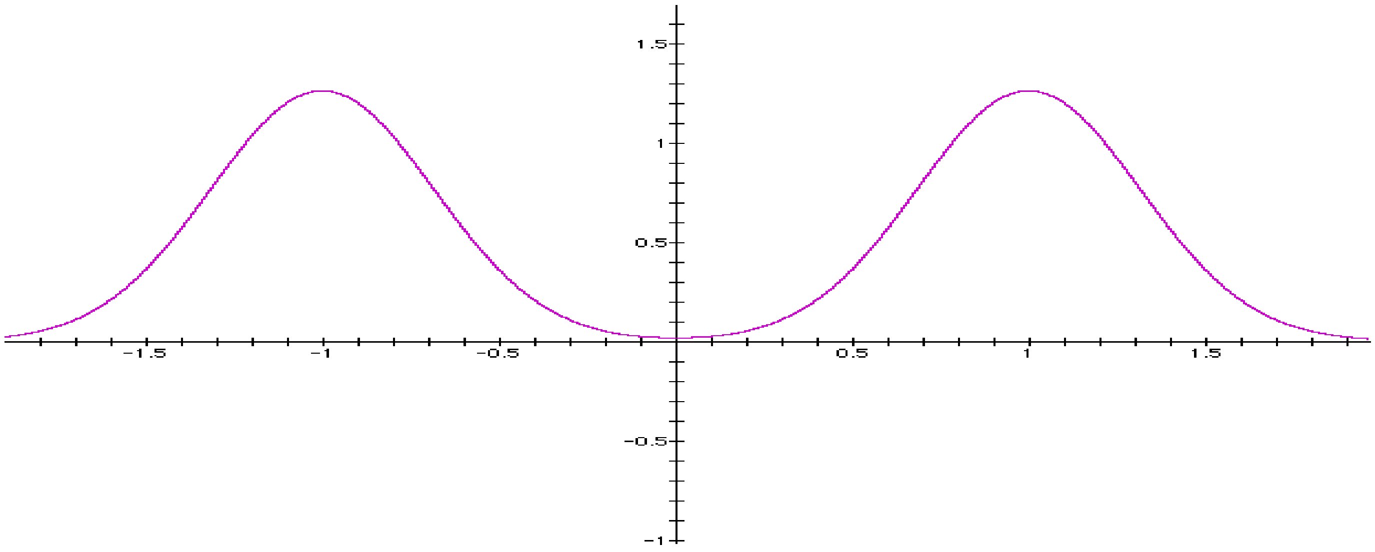
- Definition for discrete random variables: $E\{\mathbf{x}\} = \sum_i \mathbf{x}_i P(\mathbf{x}_i) = \langle \mathbf{x} \rangle$
- Definition for continuous random variables: $E\{\mathbf{x}\} = \int_{-\infty}^{+\infty} \mathbf{x}_i p(\mathbf{x}_i) d\mathbf{x} = \langle \mathbf{x} \rangle$
- E{x} is often called the MEAN of x.
- E{x} is the "Center of Mass" of the distribution.

  - Example I: What is the mean of a normal distribution?

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right)$$

Expected Values (cont'd)

- Example II: What is the mean of the distribution below?



Note: The Expectation of a variable is often assumed to be the most probable value of the variable -- but this may go wrong!

Sample Expectation

- Given a FINITE sample of data, what is the Expectation?

$$E\{\mathbf{x}\} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$$

Expectation of Function of Random Variables

$$E\{g(\mathbf{x})\} = ?$$

- as long as sum (or integral) remain bounded, just replace x*p(x) with g(x)*p(x) in E{}

- **Note: in general,** $E\{g(\mathbf{x})\} \neq g(E\{\mathbf{x}\})$
- **Other rules:**

$$E\{a\mathbf{x}\} = aE\{\mathbf{x}\}$$

$$E\{\mathbf{x}+\mathbf{y}\} = E\{\mathbf{x}\} + E\{\mathbf{y}\}$$

$$E\left\{\sum_i a_i \mathbf{x}_i\right\} = \sum_i a_i E\{\mathbf{x}_i\}$$

$$In\ general,\ E\{\mathbf{x}\mathbf{y}\} \neq E\{\mathbf{x}\}E\{\mathbf{y}\}$$

Variance and Standard Deviation

- **Variance** $Var\{x\} = E\left\{(x - E\{x\})^2\right\}$
- **Standard Deviation** $Std\{x\} = \sqrt{Var\{x\}}$

  - the Var gives a measure of dispersion of the data

  - Example I: What is the variance of a normal distribution?

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

  - Example II: What is the variance of a uniform distribution x ∈ [0, r] ?

$$Var\{x\} = \frac{r^2}{12}$$

  - A most important rule (but numerically dangerous):

$$Var\{x\} = E\{x^2\} - (E\{x\})^2$$

Sample Variance and Covariance

- **Sample Variance.**

$$Var\{x\} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - E\{x\}\right)^2$$

  - Why division by (N-1)? This is to obtain an unbiased estimate of the variance.

- **Covariance.**

$$Cov\{x, y\} = E\left\{\left(x - E\{x\}\right)\left(y - E\{y\}\right)\right\}$$

- **Sample Covariance.**

$$Cov\{x, y\} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - E\{x\}\right)\left(y_i - E\{y\}\right)$$

$$Cov\{\mathbf{x}\} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{x}_i - E\{\mathbf{x}\}\right)\left(\mathbf{x}_i - E\{\mathbf{x}\}\right)^T$$

Moments of a Random Variable

- **Moments**

$$m_n = E\left\{x^n\right\}$$

- **Central Moments**

$$cm_n = E\left\{(x-\mu)^n\right\}$$

- **Useful moments**:
  - $\boldsymbol{m_1}$ =Mean
  - $\boldsymbol{cm_2}$=Variance
  - $\boldsymbol{cm_3}$=Skewness (measure of asymmetry of a distribution)
  - $\boldsymbol{cm_4}$=Kurtosis (detects heavy and light tails and deformations of a distribution; important in computer vision)

Joint Distributions

- **Joint distributions** are distributions of **several random variables**, stating the probability that event_1 AND event_2 occur simultaneously.
  - Example 1: Generic 2 dimensional joint distribution.

$$\int_{-\infty}^{\infty} p(x, y)\, dx \;\; dy = 1$$

  - Example 2: Multivariate normal distribution in vector notation.

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right)$$

- **Marginal Distributions**: Integrate out some variables (this can be computationally very expensive).

$$p(x) = \int_{-\infty}^{\infty} p(x, y)\, dy$$

Probabilistic Independence

- By definition, **independent distributions** satisfy:

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

- Knowledge about independence is VERY powerful since it simplifies the evaluation of equations a lot.
  - Example 1: Marginal distribution of independent variables.

$$p(x) = \int_{-\infty}^{\infty} p(x,y)\,dy = \int_{-\infty}^{\infty} p(x)\,p(y)\,dy$$

$$= p(x)\int_{-\infty}^{\infty} p(y)\,dy = p(x)$$

  - Example 2: The multivariate normal distribution for independent variables.

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right)$$

$$= \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}\right)$$

Conditional Distributions

- Definition:

$$P(y \mid x) = \frac{P(x, y)}{P(x)}$$

- Since conditional distributions are more “intuitive”, some people believe that joint distributions should be defined through the more atomic conditions distribution

$$P(x, y) = P(y \mid x) P(x)$$

- What does independence mean for conditional distributions?

$$P(y \mid x) = P(y)$$

- The Chain Rule of Probabilities

$$P(x_1, x_2, \ldots, x_n) = P(x_1 \mid x_2, \ldots, x_n) P(x_2 \mid x_3, \ldots, x_n) \ldots P(x_{n-1} \mid x_n) P(x_n)$$

Bayes Rule

- Definition: $P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)}$
- Because: $P(y \mid x)P(x) = P(x, y) = P(x \mid y)P(y)$
- Interpretation:
  - P(y) is the **PRIOR** knowledge about y.
  - x is new evidence to be incorporated to update my belief about y.
  - P(x|y) is the **LIKELIHOOD** of x given that y was observed.
  - Both prior and likelihood can often be generated beforehand, e.g., by histogram statistics.
  - P(x) is a normalizing factor, corresponding to the **marginal distribution** of x. Often it need not be evaluated explicitly. But it can become a great computational burden. “P(x) is an enumeration of all possible combinations in which x and y can occur”.
  - P(y|x) is the **POSTERIOR** probability of y, i.e., the belief in y after one discovered x.