Chapter 8
Statistics for Machine Learning

8.1 Representing Data

- Numeric encoding of data
- 3 types
  - Categorical, Ordinal, Numerical

- 8.1.1 Categorical (or nominal)
  - No intrinsic ordering
  - Example
    - Jobs: soldier, sailor, tinker, spy
  - Representation
    - Integer encoding: $1, 2, 3, 4$
    - 1-of-m encoding: $(1,0,0,0)$, $(0,1,0,0)$, $(0,0,1,0)$, $(0,0,0,1)$

- 8.1.2 Ordinal
  - Discrete
  - Intrinsic ordering or ranking
  - Example
    - cold, cool, warm, hot
  - Representation
    - To preserve the ordering
    - $-1,\ \ 0,\ \ 1,\ \ 2$

- 8.1.3 Numerical
  - Values that are real numbers
  - Example
    - Temperature
    - Salary

8.2 Distributions

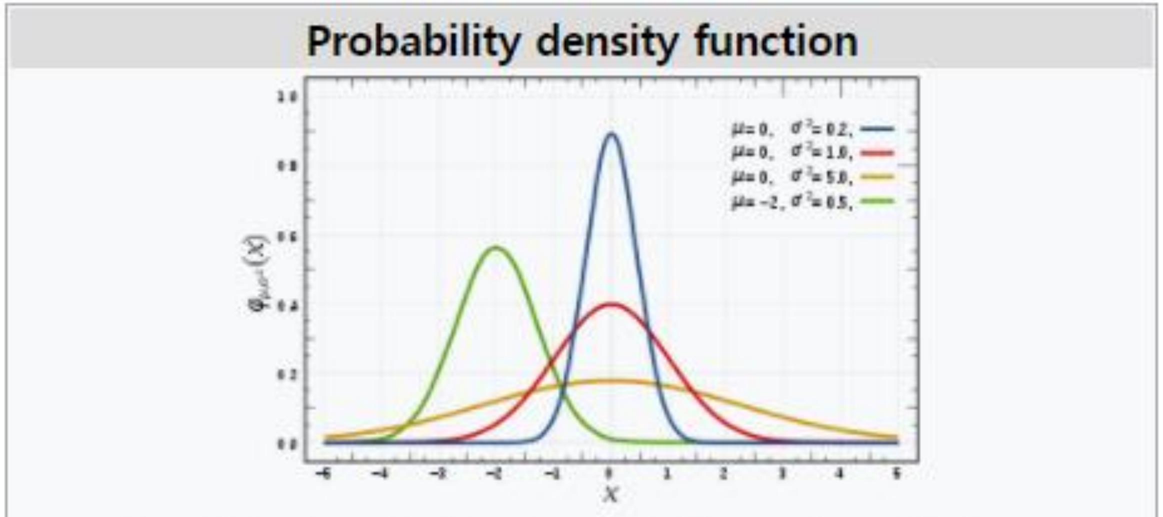
- Distributions over discrete variables
  - $\mathrm{dom}(x)$ is discrete
  - Example: dice
- Distributions over continuous variables
  - $\mathrm{dom}(x)$ is an interval or a region

- Definition 8.1 (Probability Density Functions)
  - Probability density function $p(x)$

$$p(x) \geq 0, \qquad \int_{-\infty}^{\infty} p(x)dx = 1$$

  - Probability

$$p(a \leq x \leq b) = \int_{a}^{b} p(x)dx$$



The red curve is the *standard normal distribution*

- Definition 8.2 (Averages and Expectation)
  - Average (or expectation)
    - $\langle f(x) \rangle_{p(x)}$
    - $\mathbb{E}(f(x))$

Discrete case

$$\langle f(x) \rangle_{p(x)} = \sum_{x} f(x) p(x)$$

Continuous case

$$\langle f(x) \rangle_{p(x)} = \int_{-\infty}^{\infty} f(x) p(x) dx$$

- Result 8.1 (Change of variables)

1-dimensional x

$$y = f(x)$$

$$p(x) = p(y)\frac{dy}{dx}$$

n-dimensional $\mathbf{x}$

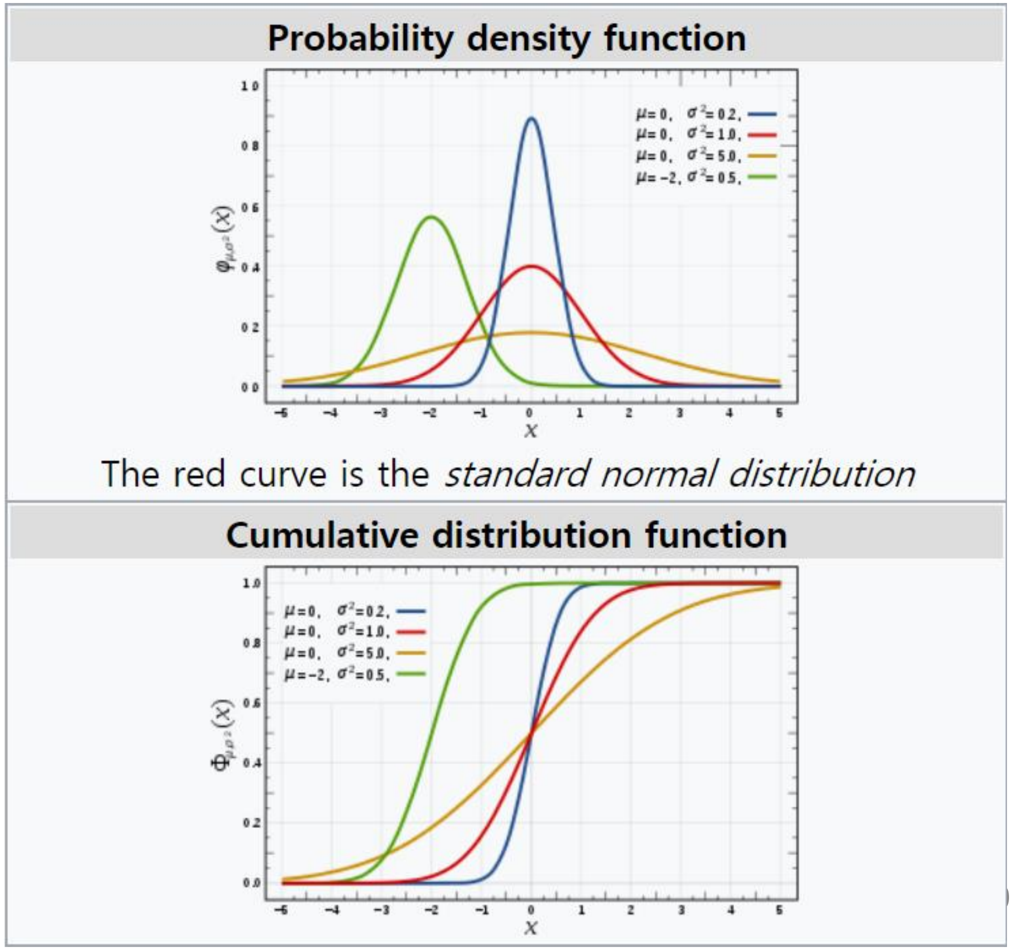
$$\mathbf{y} = f(\mathbf{x})$$

$$p(\mathbf{x}) = p(\mathbf{y})\left|\det\left(\frac{\partial\mathbf{y}}{\partial\mathbf{x}}\right)\right|$$

- Definition 8.4 (Cumulative Distribution Function)

$$\mathrm{cdf}(t) = p(x < t) = \langle \mathbb{I}[x < t] \rangle_{p(x)}$$

$$\mathrm{cdf}(-\infty) = 0,\ \mathrm{cdf}(\infty) = 1$$



- Definition 8.3 (Moments)

k-th moment

$$\left\langle x^k \right\rangle_{p(x)}$$

1st moment = mean

$$\langle x \rangle_{p(x)}$$

- Definition 8.5 (Moment Generating Function)

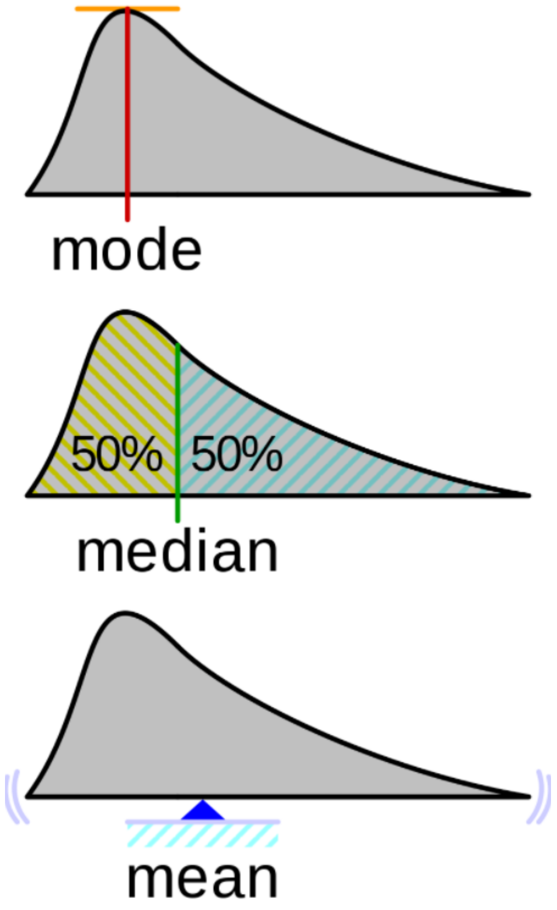
$$g(t) = \langle e^{tx} \rangle_{p(x)} = \int_{-\infty}^{\infty} e^{tx} p(x) dx$$

$$g^{(k)}(0) = \left\langle x^k \right\rangle_{p(x)}$$

- Definition 8.6 (Mode)

Highest value

$$x_* = \operatorname{argmax}_x p(x)$$



- Definition 8.7 (Variance and Correlation)

Variance

$$\sigma^2 = \langle (x - \langle x \rangle)^2 \rangle_{p(x)}$$
$$= \langle x^2 \rangle - \langle x \rangle^2$$

Covariance matrix

$$\Sigma_{ij} = \langle (x_i - \mu_i)(x_j - \mu_j) \rangle$$
$$= \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$$

Correlation matrix

$$\rho_{ij} = \left\langle \frac{x_i - \mu_i}{\sigma_i} \frac{x_j - \mu_j}{\sigma_j} \right\rangle$$

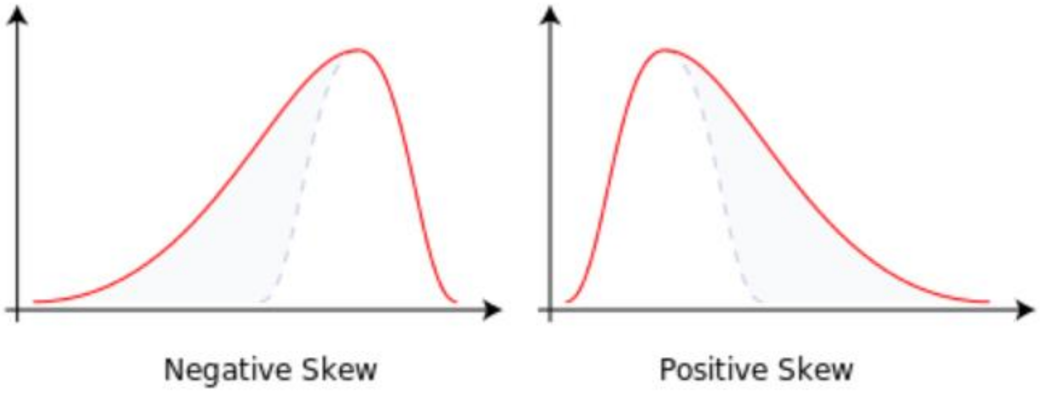
- Definition 8.8 (Skewness and Kurtosis)

Skewness

$$\gamma_1 = \frac{\left\langle (x - \langle x \rangle)^3 \right\rangle_{p(x)}}{\sigma^3}$$

Kurtosis

$$\gamma_2 = \frac{\left\langle (x - \langle x \rangle)^4 \right\rangle_{p(x)}}{\sigma^4} - 3$$



- Definition 8.10 (Empirical Distribution)

$$\text{dom}(x) = \{x^1, \dots, x^N\}$$
$$p(x) = \frac{1}{N}\sum_{n=1}^{n} \mathbb{I}[x = x^n]$$

If $x^i = x^j$ for all i and j, then
$$p(x = x^n) = \frac{1}{N}$$

Mean
$$\mu = \frac{1}{N}\sum_{n=1}^{N} x^n$$
Variance
$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N} (x^n - \mu)^2$$

- Definition 8.11 (KL divergence)

Kullback-Leibler divergence between distributions q and p

$$\mathrm{KL}(q|p) = \langle \log q(x) - \log p(x) \rangle_{q(x)}$$

KL divergence is nonnegative

$$\mathrm{KL}(q|p) \geq 0$$

Proof) ?

- Entropy

$$H(p) = -\langle \log p(x) \rangle_{p(x)}$$

$$H(p) \ = \ -\mathrm{KL}(p|u) \ + \ \mathrm{const.}$$

where u is the uniform distribution

8.3 Classical Distributions

- Definition 8.14 (Bernoulli Distribution)

$$\mathrm{dom}(x) \;=\; \{0, 1\}$$
$$p(x = 1) = \theta$$

$$\langle x \rangle = \theta$$
$$\mathrm{var}(x) = \theta(1 - \theta)$$

Example: a coin toss

- Definition 8.15 (Categorical Distribution)

$$\mathrm{dom}(x) = \{1, \dots, C\}$$
$$p(x = c) = \theta_c, \ \textstyle\sum_c \theta_c = 1$$

Example: the roll of a dice

- Definition 8.16 (Binomial Distribution)

The probability that in n Bernoulli trials $x^1, \dots, x^n$
there will be k states 1 observed

$$p(y=k|\theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

where $y = \sum_{i=1}^n \mathbb{I}\left[x^i = 1\right]$, the number of 1's

$$\langle y \rangle = n\theta,\ \mathrm{var}(y) = n\theta(1-\theta)$$

Example: coin toss n times

$$\boxed{\binom{n}{k} = \frac{n!}{k!\,(n-k)!}}$$

- Definition 8.17 (Multinomial Distribution)

n trials of a categorical distribution

Example: n rolls of a dice

- Definition 8.18 (Poisson Distribution)
  - The probability of a given number of events occurring in an interval of time

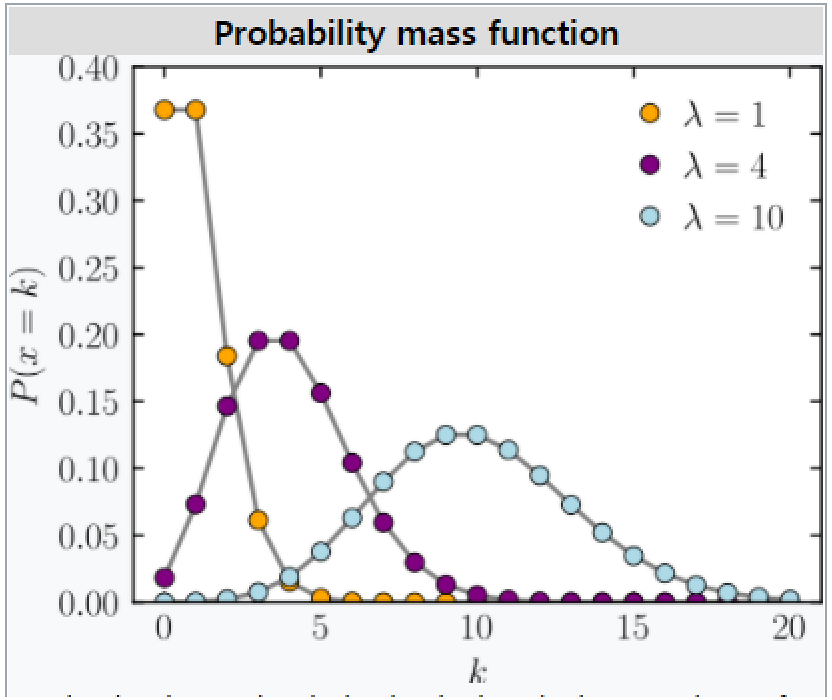
$$p(x = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

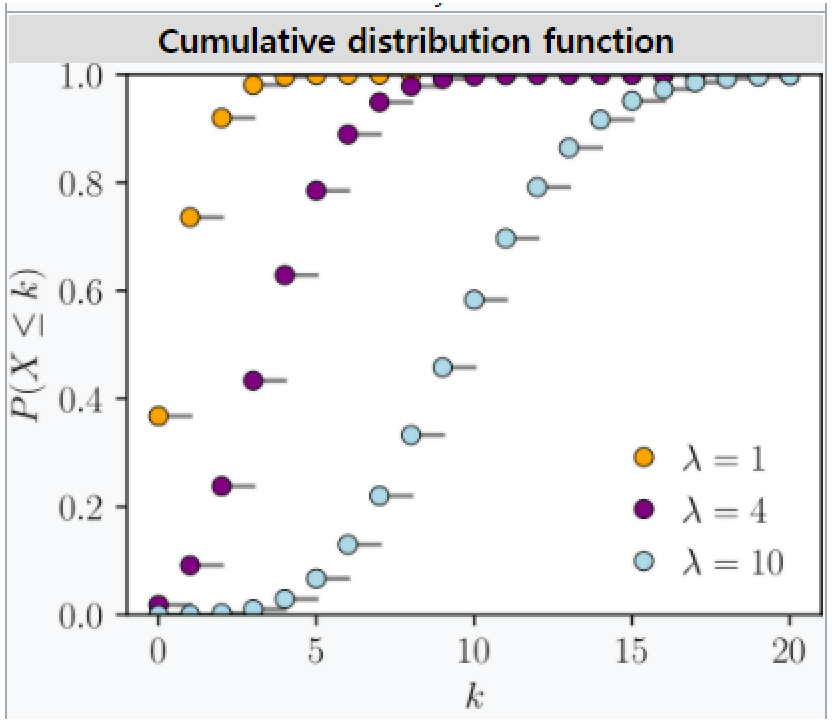
$$\boxed{\begin{gathered}\langle x \rangle = \lambda, \\ \mathrm{var}(x) = \lambda\end{gathered}}$$

  - Example: train

If λ is the expected number of events per unit interval,
then the probability of the number of events x within an interval t is

$$p(x = k|\lambda) = \frac{1}{k!} e^{-\lambda t} (\lambda t)^k$$


Probability mass function
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
$P(x = k)$
0
5
10
15
20
k
$\lambda = 1$
$\lambda = 4$
$\lambda = 10$


Cumulative distribution function
1.0
0.8
0.6
0.4
0.2
0.0
$P(X \leq k)$
0
5
10
15
20
k
$\lambda = 1$
$\lambda = 4$
$\lambda = 10$

- Definition 8.19 (Uniform distribution)

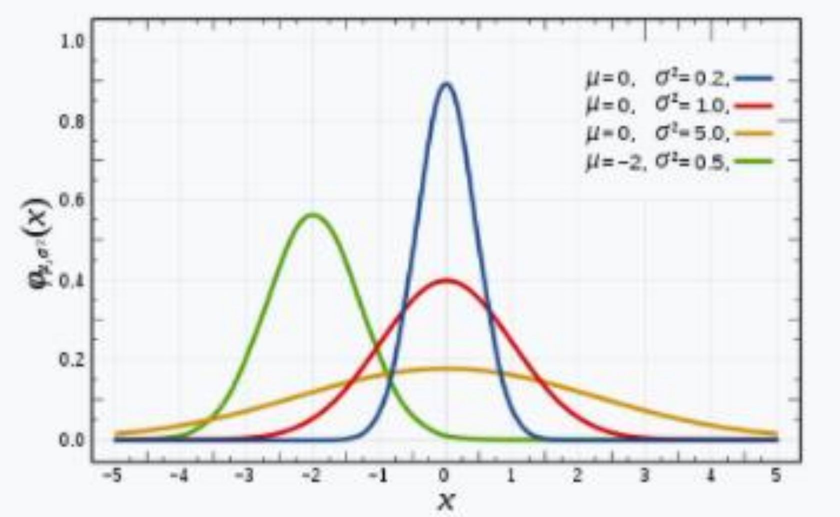
$$p(x) = \text{const.}$$

- Definition 8.21 (Gamma Distribution)
- Definition 8.22 (Inverse Gamma distribution)
- Definition 8.23 (Beta Distribution)

- Definition 8.25 (Univariate Gaussian Distribution)

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

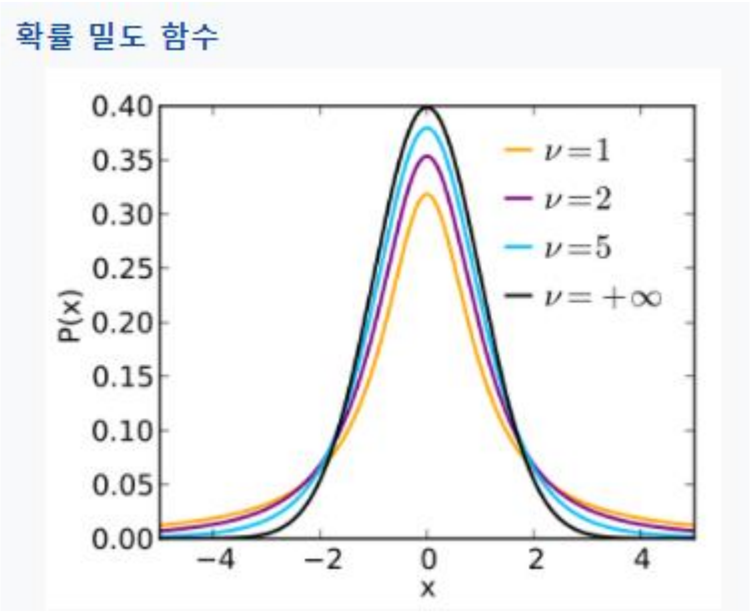
$$\langle x \rangle = \mu$$
$$\mathrm{var}(x) = \sigma^2$$

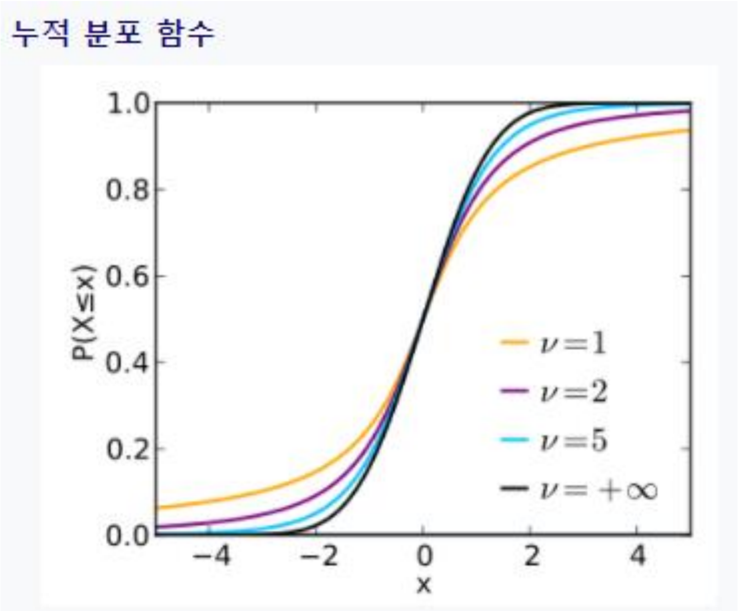


붉은 색은 표준정규분포

- Definition 8.26 (Student's t-distribution).

$$p(x|\mu,\lambda,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left(\frac{\lambda}{\nu\pi}\right)^{\frac{1}{2}}\left(1+\frac{\lambda(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

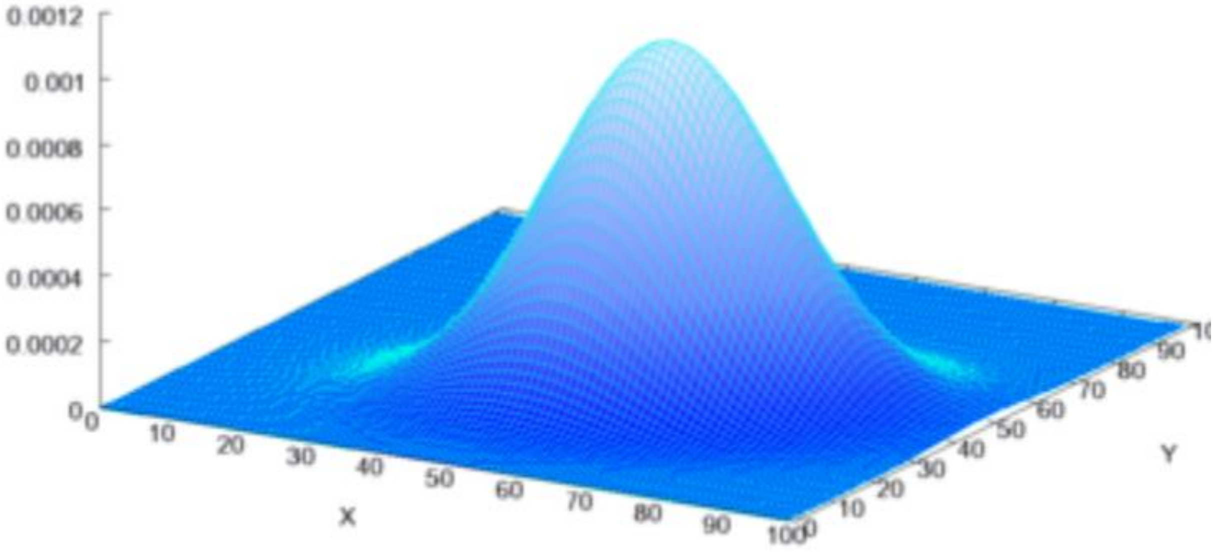




8.4 Multivariate Gaussian

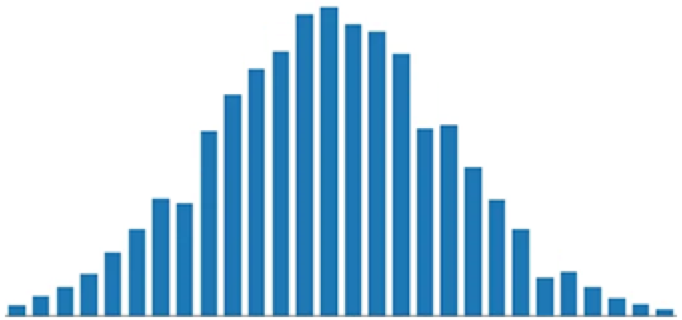
$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{\det 2\pi\boldsymbol{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

The mean vector $\boldsymbol{\mu}$
The covariance matrix $\boldsymbol{\Sigma}$



8.6 Learning distributions

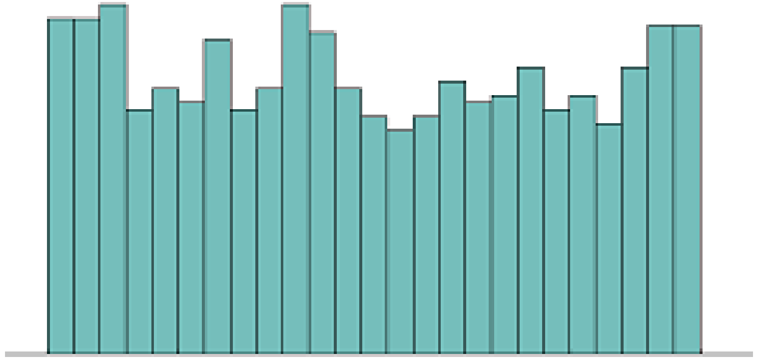
- Learning is Inferring the distribution from data
  - Inferring the distribution $p(x|\theta)$ from $\{x^1, \dots, x^N\}$
  - i.e. determining θ



Assume that the distribution is gaussian

$$\mathcal{N}(x|\mu, \sigma^2)$$

Determine μ and θ from data



Uniform distribution $p(x) = c$
Inferring is to determine c

- Maximum A posteriori
  - $\theta^{MAP} = \mathrm{argmax}_\theta \, p(\theta|\mathcal{X})$
- Maximum Likelihood
  - $\theta^{ML} = \mathrm{argmax}_\theta \, p(\mathcal{X}|\theta)$
- Moment Matching
  - θ is set such that the moment of the distribution matches the empirical moment
- Pseudo Likelihood

- Definition 8.30. Prior, Likelihood and Posterior

$$\underbrace{p(\theta|\mathcal{X})}_{\text{posterior}} = \frac{\overbrace{p(\mathcal{X}|\theta)}^{}\!\!\underbrace{}_{\text{likelihood}}\ \underbrace{p(\theta)}_{\text{prior}}}{\underbrace{p(\mathcal{X})}_{\text{evidence}}}$$

$$p(\theta|\mathcal{X}, M) = \frac{p(\mathcal{X}|\theta, M)p(\theta|M)}{p(\mathcal{X}|M)}$$

M is a model, such as gaussian

The most probable a posteriori (MAP) setting is that which maximises the posterior

$$\theta^{MAP} = \underset{\theta}{\operatorname{argmax}}\ p(\theta|\mathcal{X}, M)$$

The maximum likelihood (ML) setting is that which maximises the likelihood

$$\theta^{ML} = \underset{\theta}{\operatorname{argmax}}\ p(\mathcal{X}|\theta, M)$$

A coin is known to be $p(\mathrm{H}) = p(\mathrm{T}) = 0.5$
Experiment: toss a coin 10 times, 6 heads and 4 tails
Infer $p(\mathrm{H})$

Let $\theta = p(\mathrm{H})$ and $\mathcal{X} = \{H, H, H, H, H, H, T, T, T, T\}$
Model = the binomial distribution

ML

$$p(\mathcal{X}|\theta) = \theta^6(1-\theta)^4$$
$$\mathrm{argmax}_\theta \; p(\mathcal{X}|\theta) = 0.4$$

MAP

$$p(\theta|\mathcal{X}) = p(\mathcal{X}|\theta)p(\theta)/p(\mathcal{X})$$
$$\sim p(\mathcal{X}|\theta)p(\theta)$$

depends on the distribution of θ