Mathematics Prerequisite

Jian Tang
HEC Montreal
Mila-Quebec AI Institute
Email: jian.tang@hec.ca

Mathematics

- Linear Algebra
- Probability and Statistics
- Machine Learning Basics
- Optimization

Linear Algebra and Probability

Scalars, Vectors, and Matrices

- **Scalars**: a single value, e.g., $x = 1.5 \in R$
- **Vectors**: An array of values. A vector **x** with n dimension:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \in R^n$$

- **Matrices**: A matrix is a 2-D array of numbers, so each element is identified by two indices instead of just one

$$\boldsymbol{A} = \begin{bmatrix} A_{11}, A_{12} \\ A_{21}\ , A_{22} \end{bmatrix} \in R^{2 \times 2}$$

Transpose of Vectors and Matrices

- Transpose of a vector **x:**

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \in R^n \qquad \boldsymbol{x^T} = (x_1, x_2, \dots, x_n)$$

- Transpose a matrix $\boldsymbol{A}$: $\left(\boldsymbol{A^T}\right)_{\boldsymbol{ij}} = \boldsymbol{A_{ji}}$

$$\boldsymbol{A} = \begin{bmatrix} A_{11}, A_{12} \\ A_{21}\ , A_{22} \end{bmatrix} \qquad \boldsymbol{A^T} = \begin{bmatrix} A_{11}, A_{21} \\ A_{12}\ , A_{22} \end{bmatrix}$$

Operations

- Given two vectors: $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \in R^n$ $\quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \in R^n$

- Then $\boldsymbol{x} + \boldsymbol{y} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \dots \\ x_n + y_n \end{pmatrix}$ $\quad \boldsymbol{x} - \boldsymbol{y} = \begin{pmatrix} x_1 - y_1 \\ x_2 - y_2 \\ \dots \\ x_n - y_n \end{pmatrix}$

- Inner Product

$$\boldsymbol{x} \cdot \boldsymbol{y} = x^T y = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \sum_{k=1}^{n} x_k y_k$$

Operations

- Multiply scalar and vector

$$a \in R \qquad \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix} \in R^n \qquad a\boldsymbol{x} = \begin{pmatrix} ax_1 \\ ax_2 \\ \dots \\ ax_n \end{pmatrix} \in R^n$$

- Multiplying Matrices and Vectors: **C** = **AB**

$$\boldsymbol{C}_{ij} = \sum_k \boldsymbol{A}_{ik} \boldsymbol{B}_{kj}$$

- Note that the number of columns in **A** must be equal to the number of rows in **B**

Norms

- L^p norm of a vector $\boldsymbol{x}$

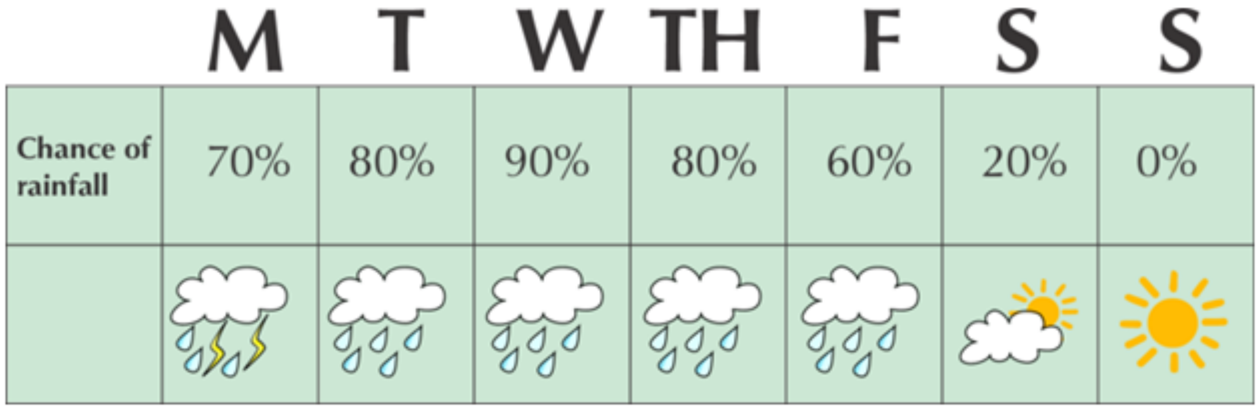
$$||\boldsymbol{x}||_{\boldsymbol{p}} = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$$

- A common one is L^2 norm

$$||\boldsymbol{x}||_{\boldsymbol{2}} = \sqrt{\sum_i x_i^2}$$

Probabilities

- Many real-world events are not certain. Probabilities are used to capture the uncertainties.
- Example:
  - What would be the outcome if I roll a dice?
  - What would be the weather like next week?



	M	T	W	TH	F	S	S
Chance of rainfall	70%	80%	90%	80%	60%	20%	0%

Random Variables & Probability Distributions

- A **random variable** is a variable that can take on different values randomly
- For example
  - X1 represents the outcome of rolling a dice $X1 \in \{1,2,3,4,5,6\}$
  - X2 represents tomorrow's weather
- A **probability distribution** is a description of how likely a random variable p(X) or a set of random variables is to take on each of its possible states p(X1, X2, …)

Discrete Random Variables and Probability Mass Functions

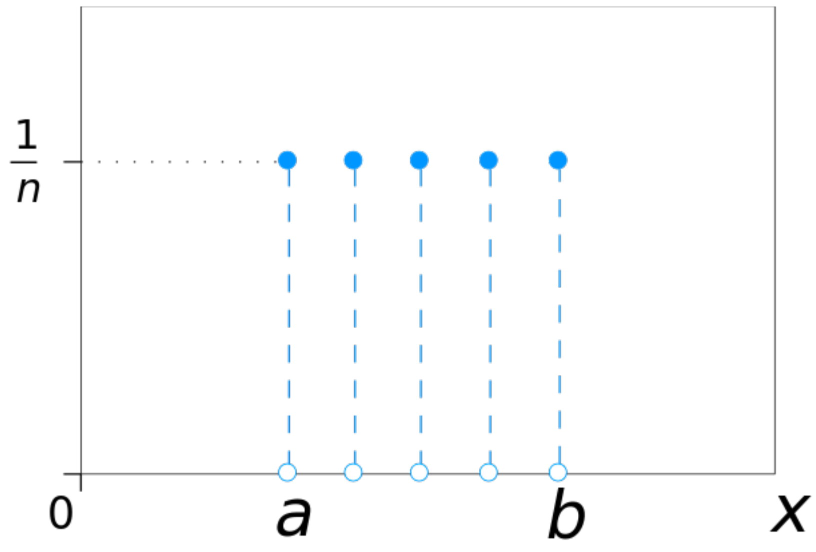
- A discrete random variable takes on a finite number of values
- A probability distribution over discrete random variables can be described using a probability mass function (PMF): $p\ (X)$

$$p(X = x_i) \geq 0, \forall i$$

$$\sum_i p(X = x_i) = 1$$

- Example: discrete uniform distribution

$$p(X = x_i) = \frac{1}{n}, \forall i$$



Continuous Random Variables and Probability Density Functions

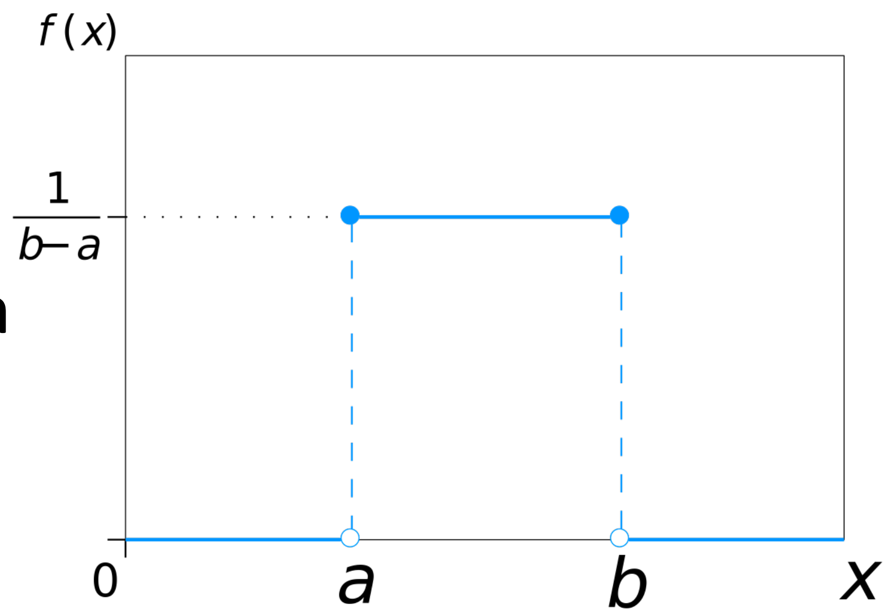
- The continuous random variables are described with probability density functions f(x):

$$f(x) \geq 0, \forall x \in X$$

$$\int f(x)dx = 1$$

- Example: continuous uniform distribution

$$f(x) = \frac{1}{b-a}, \forall a \leq x \leq b$$



Properties of Probability Distributions

- Sum rule: $p(x) = \sum_y p(x, y)$
- Product rule: $p(x, y) = p(x|y)p(y)$
- Bayes' Rule: $p(y|x) = \dfrac{p(x|y)p(y)}{p(x)}$

Expectation, Variance

- **Expectation**: the average value of X when drawn from $p\ (X)$

$$E[X] = \sum_i p(X = x_i)x_i$$

- **Variance**: a measure of how much the value x vary as we sample different values of X from its probability distribution $p\ (X)$

$$Var[X] = E\left[\left(X - E(X)\right)^2\right]$$

Binary Variable

- A Binary variable $X \in \{0, 1\}$, e.g., Flipping a coin. X = 1 representing heads and X = 0 representing tails.
- Define the probability of obtaining heads as:

$$p(X = 1) = u$$

$$p(X = 0) = 1 - u$$

$$E[X] = \mu \qquad Var[X] = \mu(1 - \mu)$$

Binomial Distribution

- The distribution of the number of observations of X=1 (e.g. the number of heads).
- The probability of observing m heads given N coin flips and a parameter μ is given by:

$$p(m\ heads|N,\mu) = Bin(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

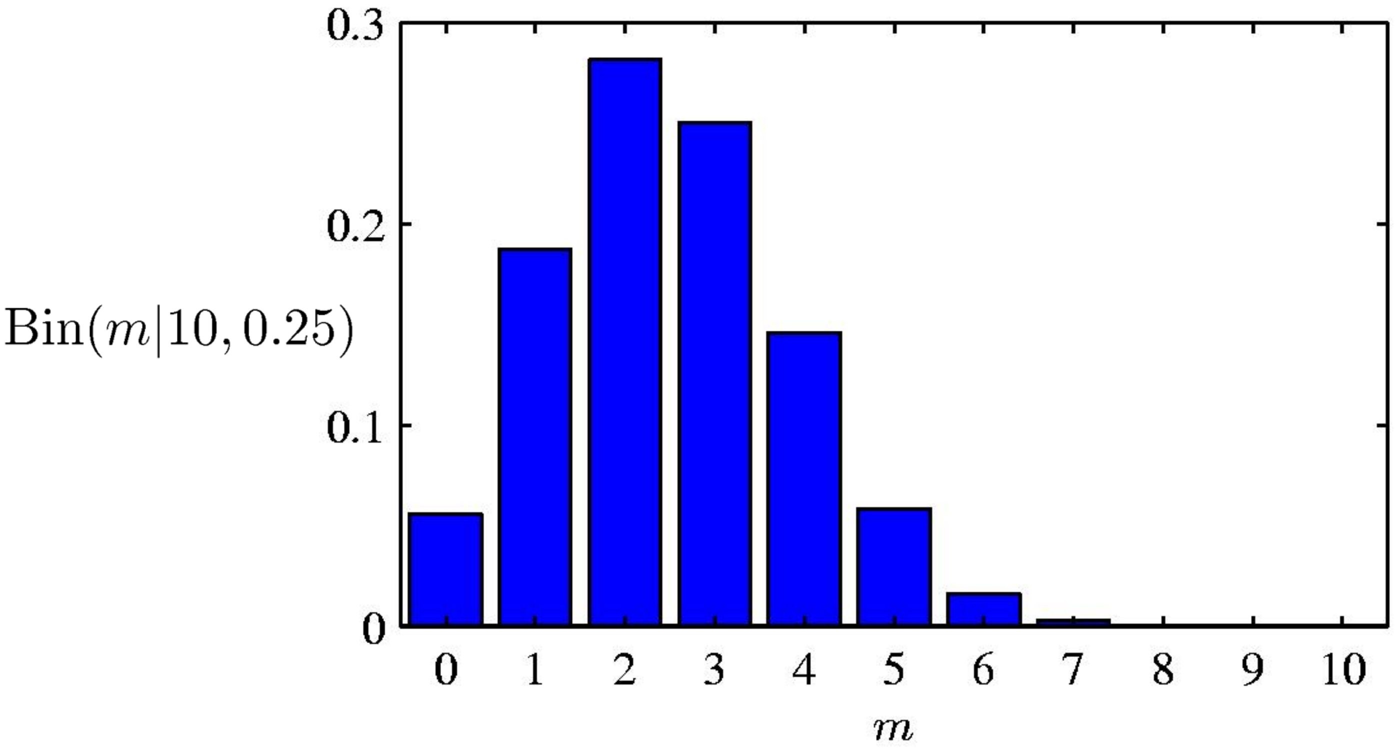
- The mean and variance can be easily derived as:

$$E[m] = \sum_{m=0}^{N} mBin(m|N,\mu) = N\mu$$

$$Var[m] = \sum_{m=0}^{N} (m-E[m])^2 Bin(m|N,\mu) = N\mu(1-\mu)$$

Example

- Histogram plot of the Binomial distribution as a function of m for N=10 and $\mu = 0.25$.



Multinomial Variables

- Consider a random variable that can take on one of K possible mutually exclusive states (e.g. roll of a dice).
- We will use so-called 1-of-K encoding scheme.
- If a random variable can take on K=6 states, and a particular observation of the variable corresponds to the state x_3=1, then **x** will be resented as:

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}}$$

- If we denote the probability of x_k=1 by the parameter μ_k, then the distribution over **x** is defined as:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \forall k : \mu_k \geqslant 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$

Multinomial Variables

- Multinomial distribution can be viewed as a generalization of Bernoulli distribution to more than two outcomes.

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

- It is easy to see that the distribution is normalized:

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1$$

- and

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}} = \boldsymbol{\mu}$$

Maximum Likelihood Estimation

- Suppose we observed a dataset $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$
- We can construct the likelihood function, which is a function of μ.

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{x_{nk}} = \prod_{k=1}^{K}\mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K}\mu_k^{m_k}$$

- Note that the likelihood function depends on the N data points only through the following K quantities:

$$m_k = \sum_n x_{nk}, \quad k = 1, ..., K.$$

- which represents the number of observations of x_k=1.
- These are called the sufficient statistics for this distribution.

Maximum Likelihood Estimation

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{x_{nk}} = \prod_{k=1}^{K}\mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K}\mu_k^{m_k}$$

- To find a maximum likelihood solution for μ, we need to maximize the log-likelihood taking into account the constraint that $\sum_k \mu_k = 1$
- Forming the Lagrangian:

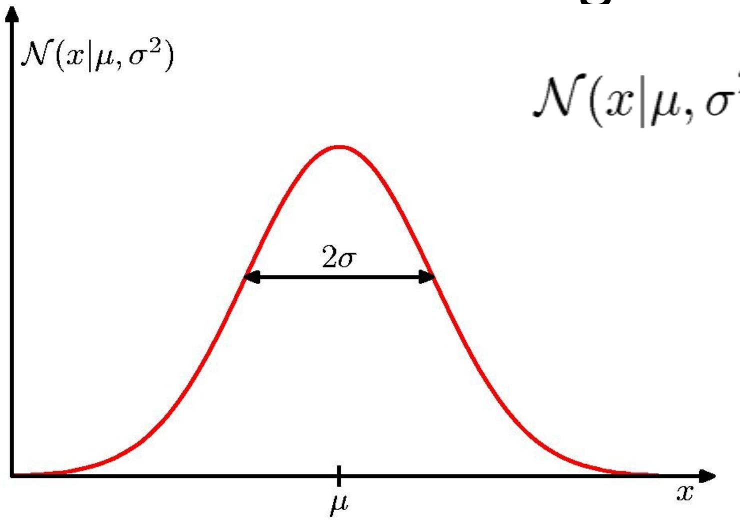
$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

$$\mu_k = -m_k/\lambda \qquad \mu_k^{\mathrm{ML}} = \frac{m_k}{N} \qquad \lambda = -N$$

which is the fraction of observations for which x_k=1.

Gaussian Univariate Distribution

- In the case of a single variable x, Gaussian distribution takes form:



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

which is governed by two parameters:

- μ (mean)
- σ^2(variance)

- The Gaussian distribution satisfies:

$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

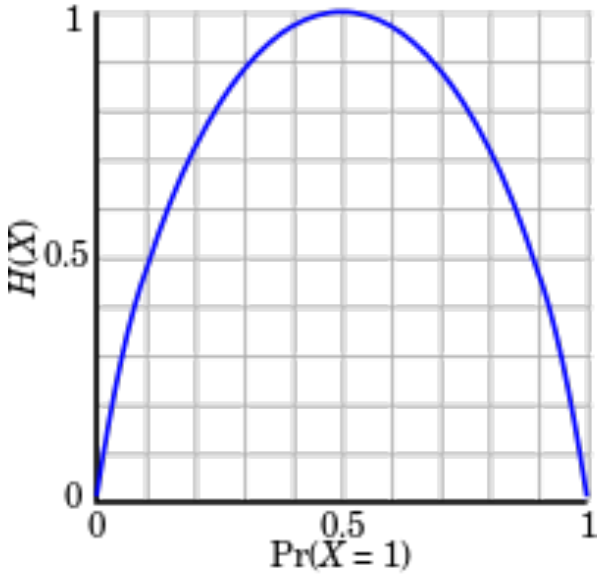
$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \mathrm{d}x = 1$$

Shannon Entropy

- The entropy H(X) of a distribution P(X) characterizes the amount of uncertainty of the random variable X.

$$H(X) = -\sum P(x) \log P(x) = -\mathbb{E}_{x \sim P} \log P(x)$$

- Example: X is a binary variable



Kullback-Leibler (KL) divergence

- KL-divergence: measure the distance between two probability distributions P(x) and Q(x)

$$D_{KL}(P||Q) = \mathbb{E}_{x\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{x\sim P}[\log P(x) - \log Q(x)]$$

- Note:
    - $D_{KL}(P||Q) \geq 0$
    - $D_{KL}(P||Q) = 0$ if and only if P=Q
    - $D_{KL}(P||Q) \neq D_{KL}(Q||P)$

Cross-Entropy H(P, Q)

- Another distance function to measure two distributions P(x) and Q(x)

$$CE(P, Q) = -\mathbb{E}_{x \sim P} \log Q(x)$$

- We can find that

$$CE(P, Q) = H(P) + D_{KL}(P||Q)$$

- Minimizing the cross-entropy with respect to Q is equivalent to minimizing the KL divergence.

Thanks!

jian.tang@hec.ca

Maximum Likelihood Estimation

- Suppose we observed i.i.d data $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$.
- We can construct the log-likelihood function, which is a function of μ and Σ:

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

- Note that the likelihood function depends on the N data points only though the following sums:

Sufficient Statistics

$$\sum_{n=1}^{N}\mathbf{x}_n \qquad \sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}$$

Maximum Likelihood Estimation

- To find a maximum likelihood estimate of the mean, we set the derivative of the log-likelihood function to zero:

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0$$

and solve to obtain:

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n.$$

- Similarly, we can find the ML estimate of Σ:

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

Maximum Likelihood Estimation

- Evaluating the expectation of the ML estimates under the true distribution, we obtain:

Unbiased estimate

$$\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] = \boldsymbol{\mu}$$

$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] = \frac{N-1}{N}\boldsymbol{\Sigma}.$$

Biased estimate

- Note that the maximum likelihood estimate of Σ is biased.
- We can correct the bias by defining a different estimator:

$$\widetilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

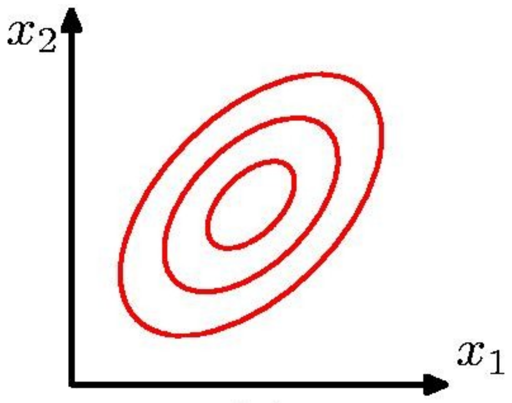
Discussion: Connections between Maximum Likelihood, KL-Divergence, and Cross Entropy

- Let P(x) be the empirical data distribution
- Let Q(x) be the distribution specified by the machine learning (a.k.a. model distribution)

Multivariate Gaussian Distribution

- For a D-dimensional vector **x**, the Gaussian distribution takes form:

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



which is governed by two parameters:

- μ is a D-dimensional mean vector.
- Σ is a D by D covariance matrix.

and $|\Sigma|$ denotes the determinant of Σ.

- Note that the covariance matrix is a symmetric positive definite matrix.