Design and Implementation of Speech Recognition Systems

Fall 2014
Ming Li

Special topic: probability basics and ML estimation
Sep 29 2014

Thanks to Felix Juefei Xu *for the contribution of the slides*

Topics To Be Covered

- Basic Probability Theory
  - Elementary Stuff
  - Bayes Rule
- Random Variables (RVs)
  - PDFs and CDFs
  - Mean and Variance
  - Commonly Used PDFs
- Joint Distributions (>1 RV)
- Conditional Probability Revisited

The Basic Stuff

- Define probability of an event as $P(A)$

$$\boxed{P(A) = \lim_{n\to\infty} \frac{n_A}{n}}$$

- Axioms of probability
  - $0 <= P(A) <= 1$
  - P (Certain Event) = 1, P (Impossible Event) =0
  - If A and B are Mutually Exclusive i.e.

$$\boxed{P[A \cap B] = 0 \text{ then } P[A \cup B] = P[A] + P[B]}$$

- A and B are Independent Events if $P(AB) = P(A)\mathrm{P}(B)$

Conditional Probability

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

Bayes Rule

$$P(B) = P(B|A_1)P(A_1) + ...P(B|A_n)P(A_n)$$

Total Probability Rule

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + ...P(B|A_n)P(A_n)}$$

Bayes Rule + Total Probability Rule

Random Variable Preliminaries

- An RV represents the probability of different events and hence takes on different values with probabilities that sum up to 1
- An RV can be Continuous, Discrete or Mixed
- Cumulative Distribution Function (CDF) – Non Decreasing Function

$$F_X(x) = P(X \leq x)$$

$$F(+\infty) = 1, F(-\infty) = 0$$

$$F(x_2) - F(x_1) = P(x_2 < x \leq x_1)$$

- Probability Density (Mass) Function (PDF or PMF)

$$f_X(x) = \frac{d}{dx}(F_X(x))$$

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

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx$$

$$F_X(x) = \sum_{x \leq x_i} f_X(x_i)$$

Mean and Variance

- Mean is also known as expected value or expectation

$$\mu = E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx$$

Continuous RV

$$\sum_{-\infty}^{+\infty} x f_X(x)$$

Discrete RV

- Variance is second moment about mean

$$\sigma^2 = E[(X - E(X)^2] = E(X^2) - E^2(X)$$

$$\int_{-\infty}^{+\infty} (x - \mu)^2 f_X(x)$$

Continuous RV

$$\sum_{-\infty}^{+\infty} (x - \mu)^2 f_X(x)$$

Discrete RV

Properties of Mean and Variance

- Expectation is a linear operator
- $E(X + c) = E(X) + E(c) = E(X) + c$
- $E(cX) = cE(X)$
- $E(X + Y) = E(X) + E(Y)$
- $E(XY) = E(X)E(Y)$ only if X and Y are uncorrelated or independent
- $\mathrm{var}(aX) = a^2\mathrm{var}(X)$

Discrete Densities

Bernoulli $\boxed{f_X(x) = x^p(1-x)^{(1-p)} \quad X = 0,1}$

Binomial $\boxed{P(X=k) = \binom{n}{k} p^k q^{n-k} \quad p+q=1 \quad k=0,1,2...n}$

Geometric $\boxed{P(X=k) = pq^{k-1} \quad k=1,2,3...\infty}$

Poisson $\boxed{P(X=k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad k=0,1,2...\infty}$

Continuous Densities

Uniform

$$f_X(x) = \frac{1}{b-a} \quad a \leq x \leq b$$
$$= 0 \quad otherwise$$

Normal

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty \leq x \leq +\infty$$

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2} dy \triangleq G(\frac{x-\mu}{\sigma})$$

$$G(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$$

Exponential

$$f_X(x) = \lambda e^{-\lambda x} \quad x \geq 0$$
$$= 0 \quad otherwise$$

Joint Distributions – Bivariate

$$\boxed{F_{X,Y}(x,y) = P(X \leq x, Y \leq y)}$$

CDF

$$\boxed{f_{X,Y}(x,y) = \frac{\delta^2 F_{X,Y}(x,y)}{\delta x \delta y}}$$

PDF

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Marginal PDFs

Joint Distributions – Bivariate

$$cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

Covariance

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Correlation Coefficient

$$E(X,Y) = E(X)E(Y)$$

Uncorrelated

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

Independent

Joint Distributions – Multivariate

$$F_{\underline{X}}(\underline{x}) = F_{X_1,X_2\ldots X_n}(x_1,x_2\ldots x_n) = P(X_1 \leq x_1 \ldots X_n \leq x_n) = \int_{-\infty}^{x_n} \ldots \int_{-\infty}^{x_1} f_{X_1,X_2\ldots X_n}(x_1,x_2\ldots x_n)dx_1\ldots dx_n$$

$$F_{\underline{X}}(-\infty \ \ldots \ -\infty) = 0 \quad F_{\underline{X}}(\infty \ \ldots \ \infty) = 1$$

CDF

$$f_{\underline{X}}(\underline{x}) = P(X_1 = x_1 \ldots X_n = x_n) = f_{X_1,X_2\ldots X_n}(x_1,x_2\ldots x_n) = \frac{dF_{\underline{X}}(\underline{x})}{d\underline{X}} = \frac{\delta^n F_{X_1,X_2\ldots X_n}(x_1,x_2\ldots x_n)}{\delta x_1 \ldots \delta x_n}$$

$$f_{\underline{X}}(\underline{x}) \geq 0$$

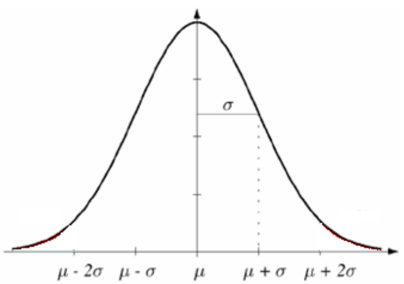
$$\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} f_{X_1,X_2\ldots X_n}(x_1,x_2\ldots x_n)dx_1\ldots dx_n = 1$$

PDF

Gaussian (Normal) Distribution

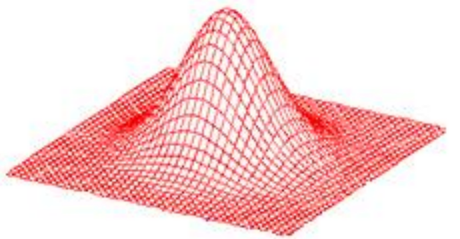
Univariate Normal Distribution

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$



Multivariate Normal Distribution

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{n/2}|\underline{\mathbf{\Sigma}}|^{1/2}} exp[\frac{-(\underline{x}-\underline{\mu})^T\underline{\mathbf{\Sigma}}^{-1}(\underline{x}-\underline{\mu})}{2}]$$



$$\text{If } \underline{X} \text{ is } N(\underline{\mu}, \underline{\mathbf{\Sigma}}) \text{ then } \underline{Y} = \underline{\mathbf{A}X} \text{ is } N(\underline{\mathbf{A}}\underline{\mu}, \underline{\mathbf{A\Sigma A}}^T)$$

Conditional Probability Revisited

$$\boxed{f_{X|Y}(x|y) = P(X = x|Y = y) = f_{X,Y}(x,y)/f_Y(y) = P(X = x, Y = y)/P(Y = y)}$$

Bayes Rule

$$\boxed{f(y|x) = f(x,y)/f(x)}$$

$$\boxed{f(x|y) = f(x,y)/f(y)}$$

$$\boxed{f(x,y) = f(x|y)f(y) = f(y|x)f(x)}$$

Simplified Notation

$$\boxed{f(x|y) = \frac{f(x,y)}{f(y)} = \frac{f(x|y)f(y)}{f(y)} = \frac{f(y|x)f(x)}{f(y)} = \frac{f(y|x)f(x)}{\int_{-\infty}^{\infty} f(y|x)f(x)dx}}$$

The Grand Scheme

References

- Useful Denitions and Results in Probability Theory - Notes By Prof. Vijaykumar Bhagavatula for Pattern Recognition

- Athanasios Papoulis, S. Unnikrishna Pillai, “Probability, Random Variables and Stochastic Processes,” TMH 4th edition, 2002

- Richard O. Duda, Peter E. Hart, David G. Stork, “Pattern Classification,” Wiley 2nd edition, 2007

- MATLAB Help

MLE Overview

- Previous lectures have shown how to develop classifiers when the underlying statistical structure is known

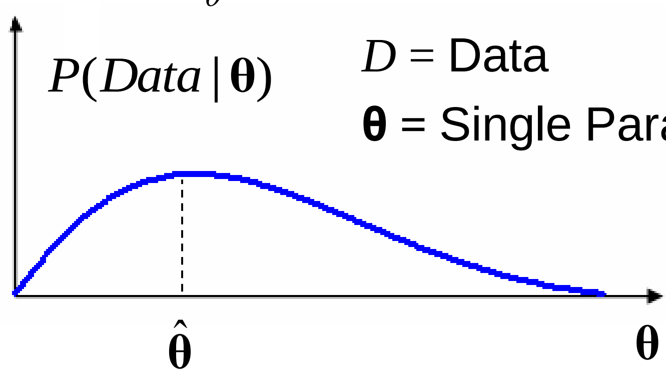
- Parametric Estimation
  - This method assumes a **particular form** of a PDF (e.g. Gaussian) is known so that we only need to determine the **parameters** (e.g. Mean & Variance)
    - Maximum Likelihood Estimation (MLE)
    - Maximum A Posteriori (Bayesian) Estimation (MAPE)

- Non-Parametric Density Estimation
  - This method **does not assume ANY knowledge** about the density
    - K-Nearest Neighbor Rule

ML Estimation (MLE)

- **Maximum Likelihood Estimation**
  - Assume $P(\mathbf{x}|\omega)$ has a known parametric form uniquely determined by the parameter vector $\boldsymbol{\theta}$
  - The parameters are assumed to be **FIXED (i.e. NON RANDOM)** but unknown
  - Suppose we have a dataset D with the samples in D having been drawn **independently** according to the probability law $P(\mathbf{x}|\omega)$
  - The MLE is the value of $\boldsymbol{\theta}$ that best explains the data and **once we know this value, we know** $P(\mathbf{x}|\omega)$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \{P(D|\boldsymbol{\theta})\}$$



D = Data

$\boldsymbol{\theta}$ = Single Parameter/Vector of Parameters

"Choose the value of $\boldsymbol{\theta}$ that is the most likely to give rise to the data we observe"

MLE

$$D = \{x_1, x_2, ..., x_N\}$$ N independent observations

$$P(D \mid \boldsymbol{\theta}) = P(x_1, x_2, ..., x_N \mid \boldsymbol{\theta}) = \prod_{k=1}^{N} P(x_k \mid \boldsymbol{\theta})$$

The likelihood of observing a particular pattern (random variable)

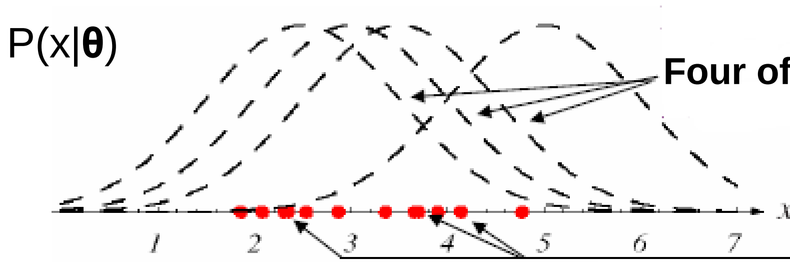
$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \{P(Data \mid \boldsymbol{\theta})\}$$

"Choose the value of θ that is the most likely to give rise to the data we observe"

MLE contd..

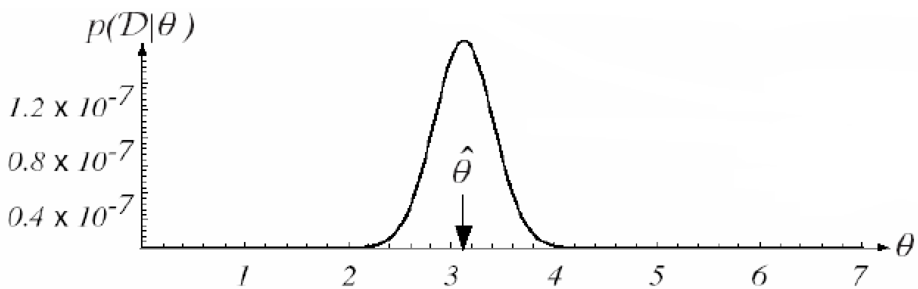
- It is convenient to work with the log of the likelihood

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \{P(D|\boldsymbol{\theta})\} = \arg\max_{\theta} \left\{\log\left(P\left(D|\boldsymbol{\theta}\right)\right)\right\}$$



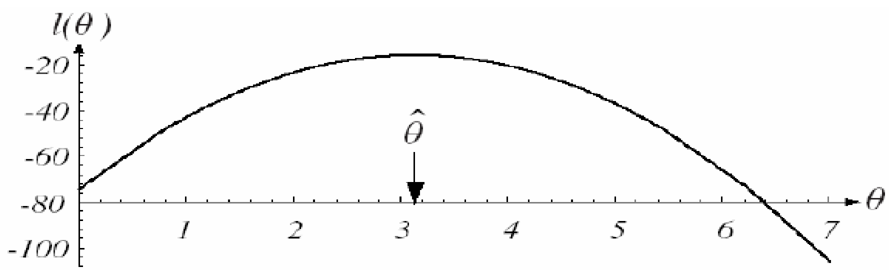
Four of the infinite number of candidate source distributions

Several training 1D points drawn from a Gaussian of a particular variance but unknown mean



The likelihood P(Data|$\boldsymbol{\theta}$) **as a function of the mean**

If we had a very large number of training points this likelihood would be very narrow



The log of the likelihood (l(θ)) is maximized at the same theta that maximizes the likelihood since log is a monotonically increasing function

How To Solve For The ML Estimate?

- Let $\boldsymbol{\theta}$ be the p-component parameter vector $\boldsymbol{\theta} = \left[\theta_1, \theta_2, \ldots, \theta_p\right]^T$
- Let this be the gradient operator $\nabla_{\boldsymbol{\theta}} = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \ldots, \frac{\partial}{\partial\theta_p}\right]^T$
- We have $P(D|\boldsymbol{\theta}) = \prod_{k=1}^{n} P\left(x_k | \boldsymbol{\theta}\right)$
- We define *l(**θ**)* the log-likelihood of the function

$$l\left(\boldsymbol{\theta}\right) = \log\left(P(D|\boldsymbol{\theta})\right) = \sum_{k=1}^{n} \log\left(P\left(x_k | \boldsymbol{\theta}\right)\right)$$

- And

$$\nabla_{\boldsymbol{\theta}} l\left(\boldsymbol{\theta}\right) = \nabla_{\boldsymbol{\theta}} \log\left(P(D|\boldsymbol{\theta})\right) = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \log\left(P\left(x_k | \boldsymbol{\theta}\right)\right)$$

- A set of necessary condition for the ML estimate can be obtained from the set of *p* equations:

$$\boxed{\nabla_{\boldsymbol{\theta}} l\left(\boldsymbol{\theta}\right) = \mathbf{0}}$$

MLE Example: Univariate Gaussian

- Now assume **neither the mean nor the covariance** matrix are known
- First consider univariate case:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 = \mu \\ \theta_2 = \sigma^2 \end{bmatrix} \quad P(D \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} P\left(x_k \mid \boldsymbol{\theta}\right) \quad \log P(x_k \mid \boldsymbol{\theta}) = -\frac{1}{2}\log\left(2\pi\theta_2\right) - \frac{1}{2\theta_2}\left(x_k - \theta_1\right)^2$$

- Its derivative is:

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \log\left(P(\mathrm{x}_k \mid \boldsymbol{\theta})\right) = \begin{bmatrix} \frac{1}{\theta_2}\left(x_k - \theta_1\right) \\ -\frac{1}{2\theta_2} + \frac{\left(x_k - \theta_1\right)^2}{2\theta_2{}^2} \end{bmatrix} = 0$$

- Setting it to zero leads to:

$$\sum_{k=1}^{n} \frac{1}{\theta_2}\left(x_k - \theta_1\right) = 0 \quad \textbf{and} \quad -\sum_{k=1}^{n} \frac{1}{\theta_2} + \sum_{k=1}^{n} \frac{(x_k - \theta_1)^2}{\theta_2^2} = 0$$

- Rearranging:

$$\boxed{\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k} \qquad \boxed{\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n} (x_k - \hat{\mu})^2}$$

ML Estimate

Sample Mean

MLE Example: Multivariate Gaussian

$$P(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}\,|\,\boldsymbol{\Sigma}\,|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$l(\theta) = \log(P(D\,|\,\boldsymbol{\theta})) = \sum_{k=1}^{n} \log(P(\mathbf{x}_k\,|\,\boldsymbol{\theta}))$$

Consider **only the mean is unknown**:

$$\log P(\mathbf{x_k}\,|\,\boldsymbol{\mu}) = -\frac{1}{2}\log\left((2\pi)^d\,|\boldsymbol{\Sigma}|\right) - \frac{1}{2}(\mathbf{x_k}-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}(\mathbf{x_k}-\boldsymbol{\mu})$$

Derivative of log likelihood must be set to 0 to obtain the MLE

$$\nabla_{\boldsymbol{\mu}} \log(P(D\,|\,\boldsymbol{\mu})) = \sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x_k}-\boldsymbol{\mu}) = \mathbf{0}$$

The ML estimate must satisfy:

$$\boxed{\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x_k}}$$

Sample Mean -> ML Estimate

MLE Example: Multivariate Gaussian

- **Neither the mean nor the covariance** matrix are known

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 = \boldsymbol{\mu} \\ \theta_2 = \boldsymbol{\Sigma} \end{bmatrix} \quad \log P(\mathbf{x_k} \mid \boldsymbol{\theta}) = -\frac{1}{2}\log\left((2\pi)^d \left|\boldsymbol{\Sigma}\right|\right) - \frac{1}{2}(\mathbf{x_k} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x_k} - \boldsymbol{\mu})$$

- Derivative of log likelihood is:

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \log\left(P(\mathrm{x_k} \mid \boldsymbol{\theta})\right) = \begin{bmatrix} \boldsymbol{\Sigma}^{-1}(\mathbf{x_k} - \boldsymbol{\mu}) \\ ? \end{bmatrix}$$

- How to take the gradient of a determinant of a matrix?

$$P(\mathbf{x} \mid \boldsymbol{\Sigma}) = \prod_{k=1}^{n} P(\mathbf{x_k} \mid \boldsymbol{\Sigma}) = \prod_{k=1}^{n} \left\{ \frac{1}{\sqrt{(2\pi)^d \left|\boldsymbol{\Sigma}\right|}} \exp\left(-\frac{1}{2}(\mathbf{x_k} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x_k} - \boldsymbol{\mu}) \right) \right\}$$

$$= \frac{1}{\left[(2\pi)^d \left|\boldsymbol{\Sigma}\right|\right]^{n/2}} \exp\left(-\frac{1}{2}\sum_{k=1}^{n}(\mathbf{x_k} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x_k} - \boldsymbol{\mu}) \right)$$

Need to take gradient with respect to Σ

MLE Example: Multivariate Gaussian

$$P(\mathbf{x}\,|\,\boldsymbol{\Sigma}) = \prod_{k=1}^{n} P(\mathbf{x_k}\,|\,\boldsymbol{\Sigma}) = \prod_{k=1}^{n}\left\{\frac{1}{\sqrt{(2\pi)^d\,|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\mathbf{x_k}-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}(\mathbf{x_k}-\boldsymbol{\mu})\right)\right\}$$

$$= \frac{1}{\left[(2\pi)^d\,|\boldsymbol{\Sigma}|\right]^{n/2}}\exp\left(-\frac{1}{2}\sum_{k=1}^{n}(\mathbf{x_k}-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}(\mathbf{x_k}-\boldsymbol{\mu})\right)$$

$$\sum_{k=1}^{n}(\mathbf{x_k}-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}(\mathbf{x_k}-\boldsymbol{\mu})$$

$$=(\mathbf{x}_1-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}(\mathbf{x}_1-\boldsymbol{\mu})+\ldots.+(\mathbf{x}_N-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})$$

$$=trace\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_1-\boldsymbol{\mu})(\mathbf{x}_1-\boldsymbol{\mu})^T\right)+\ldots+trace\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^T\right)$$

$$=trace\left(\boldsymbol{\Sigma}^{-1}(\mathbf{x}_1-\boldsymbol{\mu})(\mathbf{x}_1-\boldsymbol{\mu})^T+\ldots+\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})(\mathbf{x}_n-\boldsymbol{\mu})^T\right)$$

$$=trace\left(\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{n}(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\boldsymbol{\mu})^T\right)$$

$$\mathbf{A}=\frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\boldsymbol{\mu})(\mathbf{x}_k-\boldsymbol{\mu})^T$$

Scalar

$$\mathbf{b^TBb} = trace(\mathbf{b^TBb}) = trace(\mathbf{Bbb^T})$$

$$trace(\mathrm{A}+\mathrm{B}) = trace(\mathrm{A}) + trace(\mathrm{B})$$

$$trace(\mathbf{C(A+B)}) = trace(\mathbf{CA}) + trace(\mathbf{CB})$$

MLE Example: Multivariate Gaussian

Blackboard

Calculating derivatives against $\boldsymbol{\Sigma}^{-1}$

$$\boxed{\hat{\boldsymbol{\Sigma}}_{\mathbf{ML}} = \frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}\right)\left(\mathbf{x}_k - \hat{\boldsymbol{\mu}}\right)^T}$$

MLE - Sample Covariance