Background Knowledge Review

Le Song

Machine Learning
CS 7641,CSE/ISYE 6740, Fall 2016

Outline: Linear Algebra

- Motivating Example – Eigenfaces
- Basics
- Dot and Vector Products
- Identity, Diagonal and Orthogonal Matrices
- Trace
- Norms
- Rank and linear independence
- Range and Null Space
- Column and Row Space
- Determinant and Inverse of a matrix
- Eigenvalues and Eigenvectors
- Singular Value Decomposition
- Matrix Calculus

Motivating Example - EigenFaces

- Consider the task of representing images of faces.
- Given images of size 512x512, each image contains 262,144 dimensions or features.
- Not all dimensions are equally important in classifying faces.
- Solution – To use ideas from linear algebra, especially eigenvectors, to form a new set of reduced features.

Linear Algebra Basics - I

- Linear algebra provides a way of compactly representing and operating on sets of linear equations

  $4x_1 - 5x_2 = -13 \qquad -2x_1 + 3x_2 = 9$

  can be written in the form of $Ax = b$

  $A = \begin{bmatrix} 4 & 5 \\ -2 & 3 \end{bmatrix} \quad b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$

- $A \in \mathbb{R}^{m \times n}$ denotes a matrix with m rows and n columns, where elements belong to real numbers.

- $x \in \mathbb{R}^n$ denotes a vector with n real entries. By convention an n dimensional vector is often thought as a matrix with n rows and 1 column.

Linear Algebra Basics - II

- Transpose of a matrix results from flipping the rows and columns. Given A $\in \mathbb{R}^{m \times n}$, transpose is $A^{\top} \in \mathbb{R}^{n \times m}$

- For each element of the matrix, the transpose can be written as → $\mathrm{A}^{\top}{}_{ij} = A_{ji}$

- The following properties of the transposes are easily verified
  - $(A^{\top})^{\top} = A$
  - $(AB)^{\top} = B^{\top} A^{\top}$
  - $(A + B)^{\top} = A^{\top} + B^{\top}$

- A square matrix $A \in \mathbb{R}^{n \times n}$ is symmetric if $A = A^{\top}$.

Vector and Matrix Multiplication - I

- The product of two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ is given by $C \in \mathbb{R}^{m\times p}$, where $C_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}$

- Given two vectors $x, y \in \mathbb{R}^n$, the term $x^\top y$ (also $x \cdot y$) is called the **inner product** or **dot product** of the vectors, and is a real number given by $\sum_{i=1}^{n} x_i y_i$. For example,

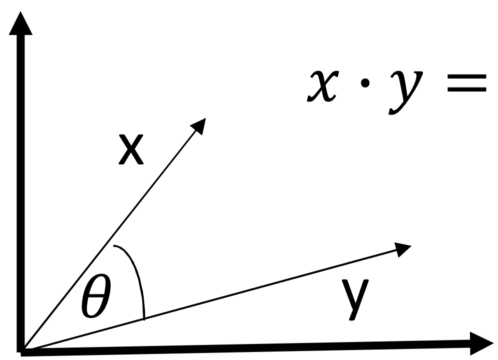
$$x^\top y = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \sum_{i=1}^{3} x_i y_i$$

- Given two vectors $x \in \mathbb{R}^n, y \in \mathbb{R}^m$, the term $xy^\top$ is called the **outer product** of the vectors, and is a matrix given by $(x_i y_j)^\top = x_i y_j$. For example,

Vector and Matrix Multiplication - II

$$xy^{\top} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & x_1y_3 \\ x_2y_1 & x_2y_2 & x_2y_3 \\ x_3y_1 & x_3y_2 & x_3y_3 \end{bmatrix}$$

- The dot product also has a geometrical interpretation, for vectors in $x, y \in \mathbb{R}^2$ with angle θ between them



$$x \cdot y = |x||y| \cos\theta$$

which leads to use of dot product for testing orthogonality, getting the Euclidean norm of a vector, and scalar projections.

Trace of a Matrix

- The trace of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted as $\boldsymbol{tr(A)}$, is the sum of the diagonal elements in the matrix
$$tr(A) = \sum_{i=1}^{n} A_{ii}$$

- The trace has the following properties
  - For $A \in \mathbb{R}^{n \times n}, tr(A) = trA^{\top}$
  - For $A, B \in \mathbb{R}^{n \times n}, tr(A + B) = tr(A) + tr(B)$
  - For $A \in \mathbb{R}^{n \times n}$, t $\in \mathbb{R}$, $tr(tA) = t \cdot tr(A)$
  - For A, B, C such that ABC is a square matrix $tr(ABC) = tr(BCA) = tr(CAB)$

- The trace of a matrix helps us easily compute norms and eigenvalues of matrices as we will see later

Linear Independence and Rank

- A set of vectors $\{x_1,\ x_2, \ldots, x_n\} \subset \mathbb{R}^m$ are said to be ***(linearly) independent*** if no vector can be represented as a linear combination of the remaining vectors. That is if

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i$$

for some scalar values $\alpha_1, \alpha_2, \ldots \in \mathbb{R}$ then we say that the vectors are linearly ***dependent***; otherwise the vectors are linearly independent

- The **column rank** of a matrix $A \in \mathbb{R}^{m \times n}$ is the size of the largest subset of columns of A that constitute a linearly independent set. **Row rank** of a matrix is defined similarly for rows of a matrix.

Range and Null Space

- The span of a set of vectors $\{x_1,\ x_2, ..., x_n\}$ is the set of all vectors that can be expressed as a linear combination of the set $\{v: v = \sum_{i=1}^{n} \alpha_i x_i\ , \alpha_i\ \in\ \mathbb{R}\}$

- If $\{x_1,\ x_2, ..., x_n\} \in \mathbb{R}^n$ is a set of linearly independent set of vectors, then span($\{x_1,\ x_2, ..., x_n\}$) = $\mathbb{R}^n$

- The range of a matrix $A \in \mathbb{R}^{m \times n}$, denoted as $\mathcal{R}(A)$, is the span of the columns of A

- The nullspace of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{N}(A)$, is the set of all vectors that equal 0 when multiplied by A
  - $\mathcal{N}(A) = \{x\ \in\ \mathbb{R}^n : Ax = 0\}$

Column and Row Space

- The row space and column space are the linear subspaces generated by row and column vectors of a matrix
- Linear subspace, is a vector space that is a subset of some other higher dimension vector space
- For a matrix $A \in \mathbb{R}^{m \times n}$
  - $Col\ space(A) = span(columns\ of\ A)$
  - $Rank(A) = \dim(row\ space(A)) = \dim(col\ space(A))$

Norms – I

- Norm of a vector $||x||$ is informally a measure of the "length" of a vector

- More formally, a norm is any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ that satisfies
  - For all $x \in \mathbb{R}^n$, $f(x) \geq 0$ (non-negativity)
  - $f(x) = 0$ is and only if $x = 0$ (definiteness)
  - For $x \in \mathbb{R}^n$, $t \in \mathbb{R}$, $f(tx) = |t| f(x)$ (homogeneity)
  - For all $x, y \in \mathbb{R}^n$, $f(x + y) \leq f(x) + f(y)$ (triangle inequality)

- Common norms used in machine learning are
  - ℓ_2 norm
    - $||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$

Norm - II

- ℓ_1 norm
  - $\|x\|_1 = \sum_{i=1}^{n} |x_i|$
- ℓ_∞ norm
  - $\|x\|_\infty = max_i |x_i|$

- All norms presented so far are examples of the family of ℓ_p norms, which are parameterized by a real number $p \geq 1$
  - $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$
- Norms can be defined for matrices, such as the Frobenius norm.
  - $\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} {A_{ij}}^2} = \sqrt{tr(A^\top A)}$

Identity, Diagonal and Orthogonal Matrices

- The identity matrix, denoted by I ∈ $\mathbb{R}^{n \times n}$ is a square matrix with ones on the diagonal and zeros everywhere else

- A diagonal matrix is matrix where all non-diagonal matrices are 0. This is typically denoted as D = $diag(d_1, d_2, d_3, \ldots, d_n)$

- Two vectors x, y ∈ $\mathbb{R}^n$ are orthogonal if $x.y$ = 0. A square matrix U ∈ $\mathbb{R}^{n \times n}$ is orthogonal if all its columns are orthogonal to each other and are normalized

- It follows from orthogonality and normality that
  - $U^T U = I = U U^T$
  - $||Ux||_2 = \ ||x||_2$

Determinant and Inverse of a Matrix

- The determinant of a square matrix $A \in \mathbb{R}^{n\times n}$ is a function *f:* $\mathbb{R}^{n\times n} \rightarrow \mathbb{R}$, denoted by $|A|$ or $\det A$, and is calculated as

  $|A| = \sum_{i=1}^{n}(-1)^{i+j}\, a_{ij}\, |A_{\backslash i, \backslash j}|$ (for any $j \in 1,2,\dots,n$)

- The inverse of a square matrix $A \in \mathbb{R}^{n\times n}$ is denoted A^{-1} and is the unique matrix such that $A^{-1}A = I = AA^{-1}$

- For some square matrices A^{-1} may not exist, and we say that A is **singular or non-invertible.** In order for A to have an inverse, A must be **full rank.**

- For non-square matrices the inverse, denoted by A^{+} ,is given by $A^{+} = (A^{\top}A)^{-1}A^{\top}$ called the **pseudo inverse**

Eigenvalues and Eigenvectors - I

- Given a square matrix $A \in \mathbb{R}^{n \times n}$ we say that $\lambda \in \mathbb{C}$ is an eigenvalue of A and $x \in \mathbb{C}^n$ is an eigenvector if

$$Ax = \lambda x, \qquad x \neq 0$$

- Intuitively this means that upon multiplying the matrix A with a vector x, we get the same vector, but scaled by a parameter λ

- Geometrically, we are transforming the matrix A from its original orthonormal basis/co-ordinates to a new set of orthonormal basis x with magnitude as λ

Eigenvalues and Eigenvectors - III

- All the eigenvectors can be written together as $AX = X\Lambda$ where the diagonals of X are the eigenvectors of A, and Λ is a diagonal matrix whose elements are eigenvalues of A

- If the eigenvector matrix X of A are invertible, then $A = X\Lambda X^{-1}$

- There are several properties of eigenvalues and eigenvectors
  - $Tr(A) = \sum_{i=1}^{n} \lambda_i$
  - $|A| = \prod_{i=1}^{n} \lambda_i$
  - Rank of A is the number of non-zero eigenvalues of A
  - If A is non-singular then $1/\lambda_i$ are the eigenvalues of A^{-1}
  - The eigenvalues of a diagonal matrix are the diagonal elements of the matrix itself!

Eigenvalues and Eigenvectors - IV

- For a symmetric matrix A, it can be shown that eigenvalues are real and the eigenvectors are orthonormal. Thus it can be represented as $U\Lambda U^\top$
- Considering quadratic form of A,
  - $x^\top Ax = x^\top U\Lambda U^\top x = y^\top \Lambda y = \sum_{i=1}^{n} \lambda_i {y_i}^2$ (where $y = U^\top x$)
- Since ${y_i}^2$ is always positive the sign of the expression always depends on λ_i. If λ_i >0 then the matrix A is positive definite, if $\lambda_i \geq 0$ then the matrix A is positive semidefinite
- For a multivariate Gaussian, the variances of *x* and *y* do not fully describe the distribution. The eigenvectors of this covariance matrix capture the directions of highest variance and eigenvalues the variance

Singular Value Decomposition

- Singular value decomposition, known as SVD, is a factorization of a real matrix with applications in calculating pseudo-inverse, rank, solving linear equations, and many others.

- For a matrix $M \in \mathbb{R}^{m \times n}$ assume $n \leq m$
  - $M = U\Sigma V^\top$ where $U \in \mathbb{R}^{m \times m}$, $V^\top \in \mathbb{R}^{n \times n}$, $\Sigma \in \mathbb{R}^{m \times n}$

  - The m columns of U, and the n columns of V are called the left and right singular vectors of M. The diagonal elements of Σ, Σ_{ii} are known as the singular values of M.

  - Let v be the i^{th} column of V, and u be the i^{th} column of U, and σ be the i^{th} diagonal element of Σ

$$Mv = \sigma u \quad \text{and} \quad M^\top u = \sigma v$$

Singular Value Decomposition - II

- $M = [u_1 \quad u_2 \; \dots \quad u_n] \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{m1} & \cdots & \Sigma_{mn} \end{bmatrix} [v_1 \quad v_2 \; \dots \quad v_n]^\top$

principal directions

Scaling factor

Projection in principal directions

- Singular value decomposition is related to eigenvalue decomposition
  - Suppose $X = [x_1 - u \quad x_2 - u \; \dots \quad x_m - u] \in \mathbb{R}^{m \times n}$
  - Then covariance matrix is $C = \frac{1}{m} X X^\top$
  - Starting from singular vector pair
    - $M^\top \mathrm{u} = \sigma v$

    $\Rightarrow M M^\top \mathrm{u} = \sigma M v$

    $\Rightarrow M M^\top \mathrm{u} = \sigma^2 u$

    $\Rightarrow C u = \lambda u$

Matrix Calculus

- For a vector x, b $\in \mathbb{R}^n$, let $f(x) = b^\top x$, then $\nabla_x b^\top x$ is equal to b
  - $\frac{\partial f(x)}{\partial x_k}$ = $\frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k$
- Now for a quadratic function, $f(x) = x^\top Ax$, with $A \in \mathbb{S}^n$, $\frac{\partial f(x)}{\partial x}$ = $2Ax$
  - $\frac{\partial f(x)}{\partial x_k}$ = $\frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{i=1}^{n} A_{ij} x_i x_j$
  - $= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2A_{kk} x_k$
  - = $2 \sum_{i=1}^{n} A_{ki} x_i$
- Let $f(X) = X^{-1}$, then $\partial(X^{-1}) = -X^{-1}(\partial X)X^{-1}$

References for self study

- Resources for review of material -
  - Linear Algebra Review and Reference by Zico Kotler
  - Matrix Cookbook by KB Peterson

Outline: Probability & Statistics

- Random variables
- Probability mass function
- Probability density function
- Cumulative distribution function
- Mean, variance, moments
- Conditional probability/density
- Independence
- Gaussian distribution
- Operations on Gaussian random variables
- Maximum likelihood estimation

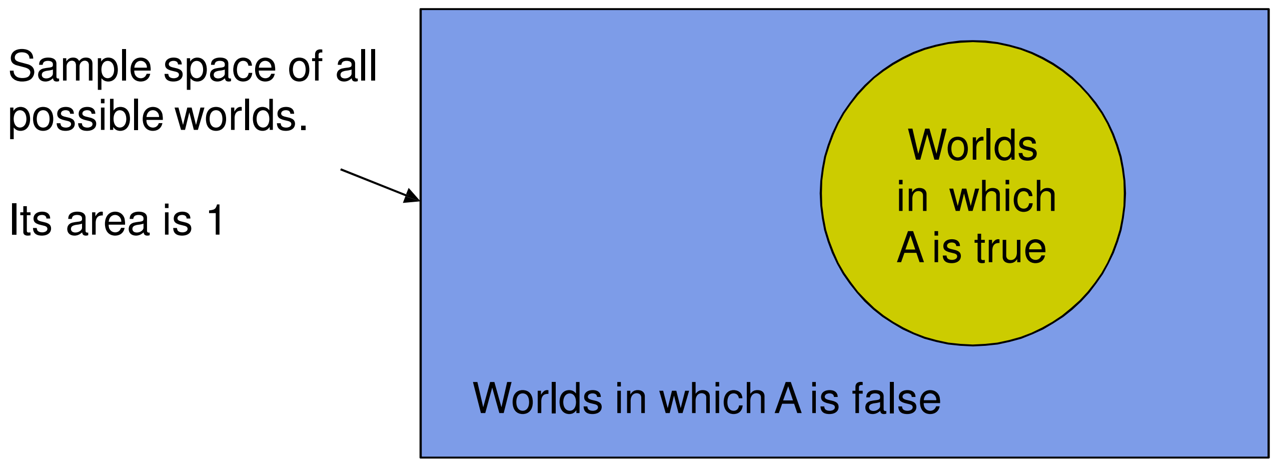
Basic Probability Concepts

- A sample space S is the set of all possible outcomes of a conceptual or physical, repeatable experiment. (S can be finite or infinite.)
  - E.g., S may be the set of all possible outcomes of a dice roll: S $(1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6)$
  - E.g., S may be the set of all possible nucleotides of a DNA site: S $(A \quad C \quad G \quad T)$

  - E.g., S may be the set of all possible time-space positions of an aircraft on a radar screen.
- An Event A is a set of outcomes of an experiment
  - Seeing "1" or "6" in a dice roll; observing a "G" at a site; UA007 in space-time interval

Probability

- An *event space* E is the collection of all possible events
  - All dice-rolls, reading a genome, monitoring the radar signal
- A probability P(A) is a function that maps an event A onto the interval [0,1]. P(A) is also called the probability measure or probability mass of A.



P(A) is the area of the oval

Kolmogorov Axioms

- For any event $A, B \subseteq S$:
  - $1 \geq P(A) \geq 0$
  - $P(S) = 1$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Random Variable

- A random variable is a function that associates a unique numerical value (a token) with every outcome of an experiment. (The value of the r.v. will vary the experiment is repeated)
- RV Distributions:
  - Continuous RV:
    - The outcome of observing the measured location of an aircraft
    - The outcome of o
  - Discrete RV:
    - The outcome of a dice-roll
    - The outcome of a coin toss

Discrete Prob. Distribution

- A probability distribution P defined on a discrete sample space S is an assignment of a non-negative real number P(s) to each sample s∈S :
  - Probability Mass Function (PMF): $p_x(x_i) = \mathrm{P}[X = x_i]$
  - Properties: $p_x(x_i) \geq 0$ and $\sum_i p_X(x_i) = 1$

- Examples:
  - Bernoulli Distribution:
    - $\begin{cases} 1-p & for\ x = 0 \\ p & for\ x = 1 \end{cases}$
  - Binomial Distribution:
    - P$(X = k)$=$\binom{n}{k} p^k (1-p)^{n-k}$

Continuous Prob. Distribution

- A continuous random variable X is defined on a continuous sample space: an interval on the real line, a region in a high dimensional space, etc.
  - It is meaningless to talk about the probability of the random variable assuming a particular value --- P(x) = 0
  - Instead, we talk about the probability of the random variable assuming a value within a given interval, or half interval, or arbitrary Boolean combination of basic propositions.
  - Cumulative Distribution Function (CDF): $F_x(x) = \mathrm{P}[X \leq x]$
  - Probability Density Function (PDF): $F_x(x) = \int_{-\infty}^{x} f_x(x)\, dx$ or $f_x(x) = \frac{d\, F_x(x)}{dx}$
  - Properties: $f_x(x) \geq 0$ and $\int_{-\infty}^{\infty} f_x(x)\, dx = 1$
  - Interpretation: $f_x(x) = \mathrm{P}[X \in \frac{[x, x+\Delta]}{\Delta}]$

Continuous Prob. Distribution

- Examples:
  - Uniform Density Function:

$$f_x(x) = \begin{cases} \dfrac{1}{b-a} & for\ a \le x \le b \\ 0 & otherwise \end{cases}$$

  - Exponential Density Function:

$$f_x(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \qquad for\ x \ge 0$$

$$F_x(x) = 1 - e^{\frac{-x}{\mu}} \qquad for\ x \ge 0$$

  - Gaussian(Normal) Density Function

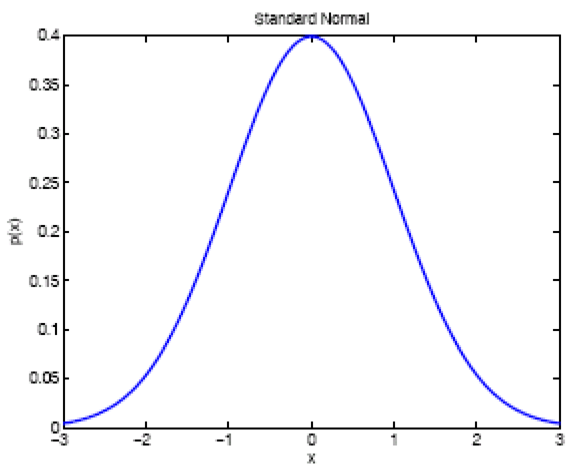
$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

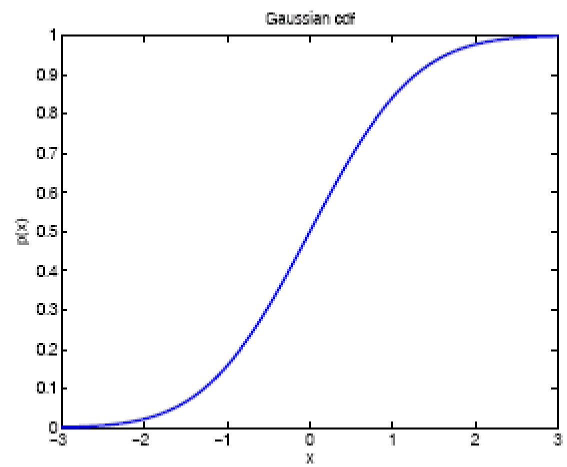
Continuous Prob. Distribution

- Gaussian Distribution:
  - If Z~N(0,1)

$$F_x(x) = \Phi(x) = \int_{-\infty}^{x} f_x(z)dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{\frac{-z^2}{2}}dz$$

  - This has no closed form expression, but is built in to most software packages (eg. normcdf in matlab stats toolbox).





Characterizations

- Expectation: The mean value, center of mass, first moment:

$$E_X[g(X)] = \int_{-\infty}^{\infty} g(x) p_X(x) dx = \mu$$

- N-th moment: $g(x) = x^n$
- N-th central moment: $g(x) = (x - \mu)^n$
- Mean: $E_X[X] = \int_{-\infty}^{\infty} x p_X(x) dx$
  - $E[\alpha X] = \alpha E[X]$
  - $E[\alpha + X] = \alpha + E[X]$
- Variance(Second central moment): $Var(x) = E_X[(X - E_X[X])^2] = E_X[X^2] - E_X[X]^2$
  - $Var(\alpha X) = \alpha^2 Var(X)$
  - $Var(\alpha + X) = Var(X)$

Joint RVs and Marginal Densities

- Joint cumulative distribution: $F_{X,Y}(x,y) = \mathrm{P}[\{X \le x\} \cap \{Y \le y\}] = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(\alpha,\beta) d\alpha d\beta$

- Marginal densities:
  - $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,\beta) d\beta$
  - $p_X(x_i) = \sum_j p_{X,Y}(x_i, y_j)$

- Expectation and Covariance:
  - $E[X+Y] = E[X] + E[Y]$
  - $cov(X,Y) = E[(X - E_X[X])(Y - E_Y(Y)] = E[XY] - E[X]E[Y]$
  - $Var(X+Y) = Var(X) + 2cov(X,Y) + Var(Y)$

Conditional Probability

- P(X|Y)= Fraction of the worlds in which X is true given that Y is also true.

- For example:
  - H="Having a headache"
  - F="Coming down with flu"
  - $P(Headche|Flu)$ = fraction of flu-inflicted worlds in which you have a headache. How to calculate?

- Definition:

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{\mathrm{P(Y|X)P(X)}}{\mathrm{P(Y)}}$$

Corollary:

$$P(X \cap Y) = P(Y|X)P(X)$$

This is called Bayes Rule

Independence

- Recall that for events E and H, the probability of E given H, written as P(E|H), is

$$P(E|H) = \frac{P(E \cap H)}{P(H)}$$

- E and H are (statistically) independent if

$$P(E \cap H) = P(E)P(H)$$

- Or equivalently

$$P(E) = P(E|H)$$

That means, the probability of E is true doesn't depend on whether H is true or not

- E and F are conditionally independent given H if

$$P(E|H,F) = P(E|H)$$

- Or equivalently

$$P(E,F|H) = P(E|H)P(F|H)$$

Multivariate Gaussian

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\}$$

- Moment Parameterization $\mu = E(X)$

$$\Sigma = Cov(X) = E[(X-\mu)(X-\mu)^{\top}]$$

- Mahalanobis Distance $\Delta^2 = (x = \mu)^{\top}\Sigma^{-1}(x-\mu)$
- Canonical Parameterization

$$p(x|\eta,\Lambda) = \exp\{a + \eta^{\top}x - \frac{1}{2}x^{\top}\Lambda x\}$$

where $\Lambda = \Sigma^{-1},\ \eta = \Sigma^{-1}\mu, a = -\frac{1}{2}(n\ log2\pi - \log|\Lambda| + \eta^{\top}\Lambda\eta\ )$

- Tons of applications (MoG, FA, PPCA, Kalman filter,…)

Multivariate Gaussian P (X1, X2)

- Joint Gaussian $P(X_1, X_2)$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

- Marginal Gaussian

$$\mu_2^m = \mu_2 \qquad \Sigma_2^m = \Sigma_2$$

- Conditional Gaussian $P(X_1 | X_2 = x_2)$

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Operations on Gaussian R.V.

- The linear transform of a Gaussian r.v. is a Gaussian. Remember that no matter how x is distributed

$$E(AX+b) = AE(X) + b$$

$$Cov(AX+b) = ACov(X)A^{\top}$$

this means that for Gaussian distributed quantities:

$$X \sim N(\mu, \Sigma) \quad \rightarrow \quad AX + b \sim N(A\mu + b, A\Sigma A^{\top})$$

- The sum of two independent Gaussian r.v. is a Gaussian

$$\mathrm{Y} = \mathrm{X}_1 + \mathrm{X}_2,\ \mathrm{X}_1 \perp X_2 \ \rightarrow \mu_y = \mu_1 + \mu_2, \Sigma_y = \Sigma_1 + \Sigma_2$$

- The multiplication of two Gaussian functions is another Gaussian function (although no longer normalized)

$$N(a, A)N(b, B) \propto N(c, C),$$
$$where\ C = (A^{-1} + B^{-1})^{-1}, c = \mathrm{C}\mathrm{A}^{-1}a + CB^{-1}b$$

MLE

- Example: toss a coin
- Objective function:

$$l(\theta; Head) = \log P(Head|\theta) = \log \theta^n (1-\theta)^{N-n} = n \log \theta + (N-n) \log(1-\theta)$$

- We need to maximize this w.r.t. θ
- Take derivatives w.r.t. θ

$$\frac{dl}{d\theta} = \frac{n}{\theta} - \frac{N-n}{1-\theta} = 0$$

$$\longrightarrow \quad \hat{\theta}_{MLE} = \frac{n}{N}$$

Central Limit Theorem

- If (X_1 ,X_2, … X_n) are i.i.d. continuous random variables, then the joint distribution is $f(\bar{X})$

- CLT proves that $f(\bar{X})$ is Gaussian with mean $E[X_i]$ and $Var[X_i]$

$$\bar{X} = f(X_1, X_2, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad as\ n \rightarrow \infty$$

- Somewhat of a justification for assuming Gaussian noise