STATISTICAL AND MATEMATICAL METHODS FOR AI:

ERRORS: discrete representations for continous numbers lead to approximation errors (measurement, arithmetic, truncation, inherent)

ABSOLUTE ERROR: let x be a ground truth and $\tilde{x}$ be the result of some operation, then $E_x := |\tilde{x} - x|$

RELATIVE ERROR: $R_x := \frac{E_x}{|x|}$, facilitates understanding the error's order of magnitude

ROUNDING RULES: rules for approximation of $x \in \mathbb{R}$ to $fl(x) \in \mathbb{F}$

Truncation: $fl(x) = x_\beta$ truncated after the β-th digit

Rounding: $fl(x) = y \in \mathbb{F}$ nearest to $x \in \mathbb{R}$, most accurate default IEEE.

FLOATING POINT SYSTEM: a floating point system is defined as $\mathbb{F} = (\beta, t, L, U)$ where β=base, t=precision, and $[L, U]$=exponent range

NORMALIZED SCIENTIFIC NOTATION: let $x \in \mathbb{R}$ s.t. $\alpha_1, \dots, \alpha_n \neq 0$ are the first n non-zero digits of x, then $fl(x) = 0,\alpha_1 \dots \alpha_n \cdot \beta^k$ is the normalized representation $\in \mathbb{F}(\beta, n, L, U)$ for x

MACHINE PRECISION: accuracy of the floating point system, $\min(\varepsilon)$ s.t. $fl(1+\varepsilon) > 1$, maximum relative representation error $(\forall x \in \mathbb{R} \ \frac{|fl(x) - x|}{|x|} < \varepsilon)$

N.B.: 1) $\varepsilon = \beta^{1-t}$ with rounding by truncation

2) $\forall$ arithmetic operation $\cdot : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, $\odot : \mathbb{F} \times \mathbb{F} \to \mathbb{F}$ causes a rounding error corresponding to $\left| \frac{(x \odot y) - (x \cdot y)}{(x \cdot y)} \right| < \varepsilon$

LINEAR SYSTEMS: $Ax = b \Leftrightarrow \exists A^{-1} (A_{\text{non singular}}) \Leftrightarrow \det(A) \neq 0 \Leftrightarrow x = A^{-1}b$, A^{-1} is an expensive computation $O(n^3)$

LU FACTORIZATION: direct method for linear systems $A\bar{x} = \bar{b}$, let $A = LU$ (lower and upper triangulars) then $A\bar{x} = \bar{b} \equiv LU\bar{x} = \bar{b}$ thus $Ly = b$ and $Ux = y$

N.B.: unstable, subject to large approximation errors in computing L and U.

PIVOTING: let $A' = PA$ where $P = I$ with swapped rows then $A\bar{x} = \bar{b} \equiv A'\bar{x} = P\bar{b}$ thus $LU\bar{x} = P\bar{b}$, $Ly = P\bar{b}$ and $Ux = y$

CHOLESKY FACTORIZATION: let $A^{n \times n} > 0$ then $A = LL^T$ and $A\bar{x} = \bar{b} \equiv LL^T\bar{x} = \bar{b}$ thus $Ly = b$ and $Lx = y$

AFFINE SPACES: let V be a vector space, $x_0 \in V$ and $U \subseteq V$ be a subspace then $L := x_0 + U := \{x_0 + u \mid u \in U\} = \{v \in V \mid \exists u \in U \text{ s.t. } v = x_0 + u\} \subseteq V$ is said to be an affine subspace (linear manifold) of V, furthermore U and x_0 are said to be direction and support point of L

N.B.: 1) if $x_0 \notin U$ then $\bar{0} \notin L$ thus $L \neq$ vector subspace of V for $x_0 \in U$.

2) affine subspaces can be described by parameters, indeed $\forall x \in L \ x = x_0 + \lambda_1 b_1 + \dots + \lambda_n b_n$ where $(b_1, \dots, b_n)$ = ordered basis of U.

AFFINE MAPPINGS: let $\Phi : V \to W$ be a linear mapping and $w \in W$, then $\Psi : V \to W$ s.t. $\Psi(x) = w + \Phi(x)$ is said to be an affine mapping

N.B.: w is said to be the translation vector of Ψ.

NORM: $\|\cdot\| : V \to \mathbb{R}$ assigns a notion of length s.t.:

1) $\forall x \in V \ \|x\| \geq 0 \wedge \|x\| = 0 \Leftrightarrow x = 0$

2) $\|\lambda \cdot x\| = |\lambda| \cdot \|x\|$

3) $\|x + y\| \leq \|x\| + \|y\|$

MANHATTAN NORM: $\|x\|_1 := \sum_{i=1}^{n} |x_i|$

EUCLIDIAN NORM: $\|x\|_2 := \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x^T \cdot x}$

N.B.: in general $\|\cdot\|_p = \left(\sum_{i=1}^{n} |x|^p\right)^{\frac{1}{p}}$

DOT PRODUCT: let $x,y \in V$, the dot product of x and y is defined as $\langle x^T, y\rangle := \sum_{i=1}^{n} x_i y_i$

MATRIX NORM: The norm of a matrix is defined as a function $\|\cdot\|: \mathbb{R}^{n\times m} \longrightarrow \mathbb{R}$ s.t.

1) $\forall A^{n\times m}\ \|A\| \geq 0 \wedge \|A\| = 0 \iff A = 0^{n\times m}$

2) $\forall \alpha \in \mathbb{R}, A^{n\times m}\ \|\alpha A\| = \alpha \|A\|$

3) $\forall A^{n\times m}, B^{n\times m}\ \|A+B\| \leq \|A\| + \|B\|$

FROBENIUS NORM: $\|A\|_F := \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2}$

2-NORM: $\|A\|_2 := \sqrt{\rho(A^TA)}$ where $\rho(A) := \max(\text{eigvals}(A))$ (=spectral radius)

POSITIVE DEFINITE MATRIX: let $A^{n\times n}$, $V \subseteq \mathbb{R}^n$ we call A positive definite if $\forall x \in V \setminus \{0\}\ x^TAx > 0$, A is positive semi-definite if $x^TAx \geq 0$

EIGENVALUES: let $A^{n\times n}$ and $\lambda \in \mathbb{C}$ be an eigenvalue of A. Then x is the corresponding eigenvector if $Ax = \lambda x$

N.B: 1) let $A^{n\times n}$ symmetric $\wedge$ eigvals(A) > 0 then $A > 0$; if eigvals(A) ≥ 0 then $A \geq 0$

2) let $A^{m\times n}$ then $A^TA \geq 0 \in \mathbb{R}^{n\times n}$ symmetric, $A^TA > 0 \iff \text{rank}(A^TA) = \min(m,n)$ (full rank)

3) let $A^{n\times m}$ then $\text{rank}(A) = n \iff \forall \lambda \in \text{eigvals}(A),\ \lambda \neq 0$

ORTHOGONAL VECTORS: let $x,y \in \mathbb{R}^n$, x and y are said to be orthogonal if $\langle x,y\rangle = 0$ $(x \perp y)$

ORTHONORMAL VECTORS: let $x,y \in \mathbb{R}^n$, x and y are said to be orthonormal if $\langle x,y\rangle = 0 \wedge \|x\| = \|y\| = 1$

ORTHOGONAL MATRICES: $A^{n\times n}$ is said to be orthogonal if it's columns are orthonormal

N.B: 1) let A be orthogonal then $AA^T = I = A^TA$, i.e $A^T = A^{-1}$

2) Transformations by orthogonal matrices don't change a vectors length

ORTHONORMAL BASIS: let $V \subseteq \mathbb{R}^n$, $\{b_1,\dots,b_m\}$ be a basis of V, if $\forall i,j \in (1,\dots,n]\ \langle b_i, b_j\rangle = \begin{cases} 0 & i \neq j \\ 1 & \text{o.w.} \end{cases}$ then $\{b_1,\dots,b_m\}$ is called an orthonormal basis.

ORTHOGONAL COMPLEMENT: let V s.t. $\dim(V) = D$ and $U \subseteq V$ s.t. $\dim(U) = M$, we call orthonormal complement of U $U^\perp \subseteq V := \{x \in V \mid \exists y \in U \text{ s.t. } x \perp y\}$ s.t. $\dim(U^\perp) = D - M$

N.B.: $U \cap U^\perp = \{0\}$ i.e $\forall x \in V\ x = \sum_{m=1}^{M} \lambda_m b_m + \sum_{j=1}^{D-M} \psi_j b_j^\perp$ where $\lambda_m, \psi_j \in \mathbb{R}$, $(b_1,\dots,b_m)$ basis of U and $(b_1^\perp,\dots,b_{D-m}^\perp)$ basis of $U^\perp$, therefore let w be the basis of $U^\perp$ we have $w \perp U$ (normal vector)

PROJECTIONS: let V be a vector space and $U \subseteq V$ a linear mapping $\pi: V \to U$ is called a projection if $\pi^2 = \pi$ $(\pi^2 = \pi \circ \pi)$

PROJECTION MATRIX: a projection matrix is defined as P_π s.t. $P_\pi^2 = P_\pi$ (=projection expressed as a matrix)

N.B: 1) let $x \in \mathbb{R}^n$, then $\pi_U(x) = u \in U$ s.t. $\|x - u\|$ is minimal, let $Ax = b$ s.t. $\nexists x$ then $\pi_U(b)$ can be used to find an approximated solution

2) $\pi_U(x) \in U$, thus $\pi_U(x) = \lambda b$ where $\lambda \in \mathbb{R}$ and b is the basis vector of U.

3) let $U \subseteq \mathbb{R}^n$ s.t. $\dim(U) = m \geq 1$ then $\pi_U(x) = \sum_{i=1}^{m} \lambda_i b_i$

ALGEBRAIC MULTIPLICITY: let $\lambda_i \in \text{eigvals}(A)$ The algebraic multiplicity of λ_i is the number of times $\sqrt{\lambda_i}$ appears in the characteristic polynomial of A

THEOREM: let $A^{n\times n}$, $(\lambda_1,\dots,\lambda_n) = \text{eigvals}(A)$ s.t. $\lambda_i \neq \lambda_j\ \forall i \neq j$ and $x_1,\dots,x_n$ be the corresponding eigenvectors then $x_1,\dots,x_n$ are linearly independent.

DEFECTIVE MATRIX: let $A^{n\times n}$ we call A defective if it has less than n linearly independent eigenvectors.

N.B.: a non-defective matrix may have less than n distinct eigenvalues.

DIAGONALIZABLE MATRIX: let $A^{n\times n}$, a is diagonalizable if $\exists P^{n\times n}$ s.t. $D=P^{-1}AP$ is diagonal, A is similar to D.

THEOREM: let $A^{n\times n}$ then A can be factored into $A=P^{-1}DP$ s.t. $P\in\mathbb{R}^{n\times n}$ and $D\in\mathbb{R}^{n\times n}$ diagonal s.t. $diag(D)=eigvals(A) \Leftrightarrow$ the eigenvectors of A are a basis of $\mathbb{R}^n$

SPECTRAL THEOREM: let $A^{n\times n}$ symmetric then $\exists$ a orthonormal basis of $V\subseteq\mathbb{R}^n$ s.t. $b_1,\dots,b_n$ are eigenvectors of A

SINGULAR VALUE DECOMPOSITION (SVD): let $A\in\mathbb{R}^{m\times n}$ rectangular $\wedge$ $rank(A)=r\in[0,\min(m,n)]$ then $\exists U\in\mathbb{R}^{m\times m}$ orthogonal, $V\in\mathbb{R}^{n\times n}$ orthogonal and $\Sigma\in\mathbb{R}^{m\times n}$ where $\Sigma_{ii}=\sigma_i\geq 0 \wedge \Sigma_{ij}=0\ \forall i\neq j$ s.t. $A=U\Sigma V^T$; $\sigma_1,\dots,\sigma_r$ are called singular values of A, by convention: $\sigma_1\geq\dots\geq\sigma_r\geq 0$

N.B.: 1) $cols(U)$ and $cols(V)$ are called respectively left and right singular vectors of A, or thonormal bases of $\mathbb{R}^m$ and $\mathbb{R}^n$.

2) We know $\forall A^{m\times n}$ rectangular $\rightarrow A^TA\geq 0$ symmetric, furthermore by the spectral theorem A^TA is diagonalizable, thus $\exists P$ orthogonal s.t. $A^TA=PDP^T$ where $D=diag(\lambda_1,\dots,\lambda_n)$

3) $\exists SVD(A)\rightarrow A^TA=(U\Sigma V^T)^T(U\Sigma V^T)=V\Sigma^TU^TU\Sigma V^T=V\Sigma^T I\Sigma V^T=V\Sigma^2V^T$, thus A^TA is similar to Σ^2 and $\sigma_i^2=\lambda_i$ where $\lambda_i\in eigvals(A^TA)$ and $\sigma_i=\sqrt{\lambda_i}$

4) $\sigma_1=\sqrt{\lambda_{max}}=\sqrt{\rho(A^TA)}=\|A\|_2$, $\sigma_r=\sqrt{\lambda_{min}}$ i.e. $\|A^n\|=1/\sigma_r\ (A^{n\times n})$

MATRIX APPROXIMATION: let $rank(A)=r$ and $A=U\Sigma V^T$ then $A_i=outer(u_iv_i^T)$ where $u_i,v_i=$ columns of U and V, $rank(A_i)=1$, $A=\sum_{i=1}^{r}\sigma_iu_iv_i^T=\sum_{i=1}^{r}\sigma_iA_i$, then $\hat{A}(k)=\sum_{i=1}^{k}\sigma_iA_i$ with $k\leq r$ is a k-rank approximation of A

N.B.: a higher rank matrix gives a better representation of the space it belongs to (rank = number of vectors forming a basis)

THEOREM: let $A,B\in\mathbb{R}^{m\times n}$ s.t. $rank(A)=r$ and $rank(B)=k$ then $\forall k<r\ \hat{A}(k)=\hat{A}(k)=\sum_{i=1}^{k}\sigma_iA_i=\arg\min_B\|A-B\|_F$, moreover $\|A-\hat{A}(k)\|_F=\sigma_{k+1}$

N.B.: if $\sigma_1,\dots,\sigma_n$ are known then k should be chosen as $\arg\max(\sigma_k-\sigma_{k+1})$ as to minimize the relative loss of information in the approximated values

PRINCIPAL COMPONENT ANALYSIS (PCA): dimensionality reduction technique, let $X\in\mathbb{R}^{d\times N}$ be a high-dimensional dataset, define $Z\in\mathbb{R}^{k\times N}$ where $k<d$ s.t. $Z=PX$ where $P=U_k^T\in\mathbb{R}^{k\times d}$

N.B.: the columns of Z are said to be the principal components of X.

BACKPROPAGATION: minimize loss function of a model through gradient descent, given a loss function L, compute ∇L w.r.t. the model's parameters:

let $y=(f_k\circ\dots\circ f_1)(x)=f_k(\dots(f(x))\dots)$ be the true result of a multi-level function and $f_i(x_{i-1})=\sigma(A_{i-1}x_{i-1}b_{i-1})$ s.t. σ is an activation function and x_{i-1} be the output of the i-th layer, then L is a measure of the distance between y and $f_k(\theta,x)$ where $\theta=\{A_0b_0,\dots,A_{k-1}b_{k-1}\}$ (= parameters), thus $\nabla L=\left(\frac{\partial L}{\partial\theta_i}\right)=\frac{\partial L}{\partial f_k}\cdot\frac{\partial f_k}{\partial f_{k-1}}\cdot\dots\cdot\frac{\partial f_i}{\partial\theta_i}\ \forall i$

AUTOMATIC DIFFERENTIATION: generalization of backpropagation, set of techniques to numerically evaluate the exact gradient of a function using intermediate variables obtained through a sequence of elementary operations resulting in automatic computation of the gradient even for complex functions

N.B.: 1) backpropagation applies automatic differentiation in the opposite direction of the data flow.

2) not all programs can be automatically differentiated

OPTIMIZATION: find optimum of a function, machine learning models can be expressed as numerical optimization methods.

OPTIMALITY CONDITIONS:

First Order: let f be differentiable around x^* local minimum then $\nabla f(x^*)=0$ (stationary point)

Second Order: 1) let f be twice differentiable around a local minimum x^* then $\nabla f(x^*)=0$ and $Hf(x^*)\geq 0$

2) let f be twice differentiable around x^* s.t. x^* is stationary $\wedge$ $Hf(x^*)>0$, then x^* is a local minimum

CONVEX SET: a set $C\subseteq\mathbb{R}^n$ is said to be convex if $\forall x,y\in C$, $\theta\in[0,1]$, $\theta x+(1-\theta)y\in C$ ($\forall$ point p between x and y $p\in C$)

CONVEX FUNCTION: let C be convex and $f:C\rightarrow\mathbb{R}$, f is said to be convex if $\forall x,y\in C\ f(\theta x+(1-\theta)y)\leq\theta f(x)+(1-\theta)f(y)$ ($\forall p$ between $f(x)$ and $f(y)$ lies above the function)

N.B.: let f be convex then $\forall$ local minimum x, x is a global minimum, $\nabla f(x)=0$ then x = global minimum

LEVEL CURVES: $L_c := \{x \in \mathbb{R}^n | f(x) = c\}$

DESCENT METHODS: solve $\min(f(x))$ by generating $x_k = x_{k-1} + \alpha_k p_k$ where $p_k \in \mathbb{R}$ is called the search direction and $\alpha_k \in \mathbb{R}$ is called the step length

Descent Direction: choose p_k s.t. $f(x_k) < f(x_{k-1})$, let f be continous and differentiable in x and $p \in \mathbb{R}^n \neq 0$ s.t. $p^T \nabla f(x) < 0$ Then p is a descent direction for f in x, $\forall f$, $-\nabla f(x)$ is a descent direction

N.B.: let p_k be a descent direction for f in x_k Then $\forall \alpha_k \in [0, \bar{\alpha}]$ $f(x_k + \alpha_k p_k) < f(x_{k-1})$ for some $\bar{\alpha}$

Step Length:

Constant: doesn't guarantee convergence

Inexact Search: choose α_k to produce sufficient decrease or a sufficient shift from $f(x_{k-1})$

Backtracking: let $\bar{\alpha} = 1$ (starting value) and $c \in]0,1]$ if $f(x_k + \bar{\alpha} p_k) < f(x_k) + c\bar{\alpha} p_k^T \nabla f(x_k)$ Then $\alpha_k = \bar{\alpha}$ else $\bar{\alpha} = \bar{\alpha} \cdot p$ (usually $p = 1/2$)

Wolfe Conditions: $f(x_k + \bar{\alpha} p_k) \leq f(x_k) + c\bar{\alpha} p_k^T \nabla f(x_k)$ (sufficient decrease, Armijo Condition), $-p_k^T \nabla f(x_k + \alpha_k p_k) \leq -c_2 p_k^T \nabla f(x_k)$ (curvature condition), satisfied by backtracking

THEOREM: if α_k is chosen by backtracking Then the descent method converges to a local minimum.

GRADIENT DESCENT: let f be continous and differentiable given a starting point x_0. $\forall k$ select $\alpha_k = c$ v by backtracking and $p_k = -\nabla f(x_k)$, iterate until conditions are met

N.B: ADAM is a gradient descent variant (different rules for α_k)

STOP CONDITIONS: $\|\nabla f(x_k)\| < \tau_1 \left(\text{or } \frac{\|\nabla f(x_k)\|}{\|\nabla f(x_0)\|} < \tau_2\right) \wedge \|x_k - x_{k-1}\| < \nu_1 \left(\text{or } \frac{\|x_k - x_{k-1}\|}{\|x_{k-1}\|} < \nu_2\right) \vee \text{iterations} > K$

STARTING GUESS: usually, random, influences result

CONVERGENCE SPEED: if $\|x_k - x^*\| < \|x_k - x^*\|^Q \cdot c$ for some c, Then the method has convergence speed $= Q$

N.B.: $Q > 0$, $Q = 1 \rightarrow$ linear speed $\wedge c < 1$; $Q \in]1,2[\rightarrow$ superlinear speed; $Q = 2 \rightarrow$ quadratic speed.

GRADIENT DESCENT WITH MOMENTUM: $x_{k+1} = x_k - \alpha_k (\nabla f(x_k))^T + \gamma \Delta x_k$ where $\Delta x_k = x_k - x_{k-1}$ and $\gamma \in [0,1]$ (x_k depends on 2 previous steps)

STOCHASTIC GRADIENT DESCENT: commonly used in neural networks, in general given N samples, $L(\theta, X, Y) = \sum_{n=1}^{N} L_n(\theta, x_n, y_n)$, Thus $\nabla_\theta L(\theta, X, Y) = \sum_{n=1}^{N} \nabla_\theta L_n(\theta, x_n, y_n)$,

let $S \subset \{1, \dots, N\}$ approximate $\nabla L(\theta, X, Y) \approx \sum_{i \in S} \nabla_\theta L_i(\theta, x_i, y_i)$, chosing a different random subset at each iteration. Let $\kappa = |S|$ Then S is said to be a minibatch of size κ

Epoch: $\neq$ iteration, an epoch ends every time the entire dataset has been used, stop condition = maximum number of epochs ($\nexists$ convergence checks, $\nexists$ backtracking)

DISCRETE DISTRIBUTIONS:

Probability Mass Function: let X be a discrete random variable, The PMF of X is a function $f_X(x): X \rightarrow [0,1]$ s.t. $f_X(x) = P(X = x)$

N.B.: 1) a PMF is a probability

2) let X be discrete $\mathbb{E}[X] = \sum_{x_i \in X} x_i f_X(x_i)$ $(= \mu)$, $\sigma^2 = \sum_{x_i \in X} (x_i - \mu)^2 \cdot f_X(x_i)$

Uniform Distribution: let X be a discrete random variable s.t. $f_X(x_i) = \frac{1}{|X|}$ $\forall x_i \in X$ Then $X \sim Unif(a, b)$ where $a = \inf(X)$ and $b = \sup(X)$

Poisson Distribution: let X be a discrete random variable and $\lambda \in \mathbb{R}$ s.t. $f_X(x_i) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$ Then $X \sim Poisson(\lambda)$

N.B: 1) The parameter λ represents the expected number of events in a given interval, The PMF represents the probability of x_i events occurring in the given interval.

2) $\mathbb{E}[X] = \sigma^2 = \lambda$

CONTINOUS DISTRIBUTIONS:

Probability Density Function: let $X:\Omega\to[a,b]$ be a continous random variable, The PDF of X is a function $f_X: X\to\mathbb{R}$ s.t $\mathbb{P}(X=x_i\in[c,d])=\int_c^d f_X(x_i)\,dx_i$

N.B: 1) $f_X(x_i)>0\ \forall x_i\in X$, $\int_X f_X(x_i)\,dx_i=1$, PDF is not a probability, its integral is ($\neq$PMF)

2) let X be continous $\mathbb{E}[X]=\int_X x_i\, f_X(x_i)\,dx_i$, $\sigma^2=\int_X (x-\mu)^2\cdot f_X(x_i)\,dx_i$, $\sigma=\sqrt{\sigma^2}$ is called standard deviation

Uniform Distribution: let X be continous in $[a,b]$ s.t. $f_X(x_i)=\frac{1}{b-a}\ \forall x_i\in X$ Then $X\sim Uni f(a,b)$

Gaussian Distribution: let X be continous s.t. $f_X(x_i)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ Then $X\sim\mathcal{N}(\mu,\sigma)$

MULTIVARIATE DISTRIBUTIONS:

Joint Probability: let X,Y be discrete, then their joint probability is given by $f_{XY}: X\times Y\to[0,1]$ s.t. $f_{XY}(x_i,y_j)=\mathbb{P}(X=x_i,Y=y_j)$

Marginal Probability: let (X,Y) be a random vector, its marginal probabilities are given by $p_x(x_i)=\mathbb{P}(X=x_i,Y)=\begin{cases}\sum_{y_j} f_{x,Y}(x_i,y_j) & Y\text{ discrete}\\ \int_{y_j} f_{XY}(x_i,y_j)\,dy_j & Y\text{ continous}\end{cases}$

CONDITIONAL PROBABILITY: $\mathbb{P}(X|Y)=\frac{\mathbb{P}(X,Y)}{\mathbb{P}(Y)}$

BAYES RULE: $\mathbb{P}(X|Y)=\frac{\mathbb{P}(Y|X)}{\mathbb{P}(Y)}\cdot\mathbb{P}(X)$

Prior: $\mathbb{P}(X)$, encapsulates subjective prior knowledge of the latent variable X

Likelihood: $\mathbb{P}(Y|X)$ describes how X and Y are related (=distribution only in y)

Posterior: $\mathbb{P}(X|Y)$, expresses what is known about X having observed Y

Evidence: $\mathbb{P}(Y)$, marginal likelihood, ensures the posterior is normalized

EXPECTED VALUE: let X be continous and $g:X\to\mathbb{R}$ then $\mathbb{E}_X[g(x)]=\int_X g(x)f_X(x)\,dx$ (Σ_X if X is discrete), for $X=(X_1,\dots,X_n)$ $\mathbb{E}_X[g(x)]=\begin{pmatrix}\mathbb{E}_{x_1}[g(x_1)]\\ \vdots\\ \mathbb{E}_{x_n}[g(x_n)]\end{pmatrix}$

COVARIANCE: let X,Y be random variables Then $\operatorname{cov}[X,Y]=\mathbb{E}_{XY}[(X-\mathbb{E}_X[X])(Y-\mathbb{E}_Y[Y])]=\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$

N.B.: for $X=(X_1,\dots,X_n)$, $Y=(Y_1,\dots,Y_n)$ $\operatorname{cov}[X,Y]=\mathbb{E}[XY^T]-\mathbb{E}[X]\mathbb{E}[Y^T]=\operatorname{cov}[X,Y]^T$

VARIANCE: let X be a random variable its variance is defined as $\operatorname{var}[X]:=\operatorname{cov}[X,X]$ ($=\sigma_X^2$)

N.B.: $\mathbb{E}[X\pm Y]=\mathbb{E}[X]\pm\mathbb{E}[Y]$, $\operatorname{var}[X\pm Y]=\operatorname{var}[X]+\operatorname{var}[Y]\pm\operatorname{cov}[X,Y]\pm\operatorname{cov}[Y,X]$

CORRELATION: $\operatorname{corr}[X,Y]=\frac{\operatorname{cov}[X,Y]}{\sqrt{\sigma^2(x)\sigma^2(Y)}}\in[-1,1]$

EMPIRICAL MEAN: let $X:\Omega\to\mathbb{R}^D$ and $x_1,\dots,x_N$ be the results of N samplings from X, Then the empirical mean of X is given by $\bar{x}:=\frac{1}{N}\sum_{i=1}^N x_i\in\mathbb{R}^D$

EMPIRICAL VARIANCE: let $X:\Omega\to\mathbb{R}^D$ and $x_1,\dots,x_N$ be the results of N samplings from X, Then the empirical covariance of X is given by $\Sigma:=\frac{1}{N}\sum_{i=1}^N(x_i-\bar{x})(x_i-\bar{x})^T$

N.B: empirical standard deviation is defined as $\sqrt{\Sigma}$

INDEPENDENCE: $X\perp\!\!\!\perp Y\iff P(X,Y)=P(X)P(Y)\iff P(Y|X)=P(Y)\iff P(X|Y)=P(X)$

N.B: $X\perp\!\!\!\perp Y\to\operatorname{var}[X+Y]=\operatorname{var}[X]+\operatorname{var}[Y]$ ($\iff\operatorname{cov}[X,Y]=0$)

CONDITIONAL INDEPENDENCE: $X\perp\!\!\!\perp Y|Z\iff P(X,Y|Z)=P(X|Z)P(Y|Z)\ \forall z\in Z$

MULTIVARIATE GAUSSIAN: let $X\in\mathbb{R}^D$ be a random vector, μ be a mean vector and Σ be a covariance matrix, if $P(X|\mu,\Sigma)=(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}\exp\left(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$ Then $X\sim\mathcal{N}(\mu,\Sigma)$

N.B: let $X,Y\sim\mathcal{N}(\mu,\Sigma)$ Then $P(X+Y)\sim\mathcal{N}(\mu_X+\mu_Y,\Sigma_X+\Sigma_Y)$

PROBABILISTIC MODELS: data = samples from $(X,Y)\sim p(X,Y|\theta)$ where θ = parameters, learn parameters that best fit the data

MAXIMUM LIKELIHOOD ESTIMATION (MLE): probabilistic parameter estimation method, define $f(\theta):=P(y|X\theta)=\prod_{n=1}^{N} P(y_n|x_n,\theta)$ (=likelihood function), $\theta^*=\operatorname{argmax}_\theta f(\theta)$

N.B.: 1) result is generally computed as $\theta^*=\operatorname{argmin}_\theta -\log(f(\theta))=L(\theta)$

2) $L(\theta)=-\log\prod_{n=1}^{N} P(y_n|x_n,\theta)=-\sum_{n=1}^{N}\log P(y_n|x_n,\theta)$

3) MLE is prone to overfitting, parameter values tend to grow large

LINEAR REGRESSION: supervised task, learn a linear function $y_i=\theta^T x_i+\varepsilon_i$ where $\varepsilon_i\sim N(0,\sigma)$

N.B.: $\varepsilon_i\sim N(0,\sigma)\iff P(y|x,\theta)\sim N(\theta^T x,\sigma)$, hence parameters can be estimated through MLE where $L(\theta)=\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i-\theta^T x_i)^2=\frac{1}{2\sigma^2}(y-X\theta)^T(y-X\theta)=\frac{1}{2\sigma^2}\|y-X\theta\|^2$ (=loss), $\nabla L(\theta):=\frac{1}{\sigma^2}X^T(X\theta-y)$

LEAST SQUARE PROBLEM: minimize $f=\|b-Ax\|^2$ where $\exists$ solutions for $Ax=b$, $\nabla f=0\iff A^TAx=A^Tb$, furthermore f is convex, thus $\operatorname{argmin}_x f=\bar{x}$ s.t. $A^TA\bar{x}=A^Tb$

N.B.: 1) linear regression can be reduced to LSP ($\operatorname{argmin}_\theta L(\theta)=\operatorname{argmin}_\theta\|y-X\theta\|^2=\theta^*$ s.t. $X^TX\theta^*=X^Ty$)

2) if $A\neq$ full rank the problem can be solved as $x=A^+b$ (pseudo-inverse)

POLYNOMIAL REGRESSION: it is possible to apply non-linear transformations to the inputs to obtain non-linear outputs: $y=\Phi_K(x)^T\theta+\varepsilon=\sum_{k=0}^{K-1}\theta_k\varphi_k(x)+\varepsilon=\sum_{k=0}^{K-1}\theta_k x^k+\varepsilon$ (K-1 degree polynomial)

N.B.: 1) let $\Phi_i=\begin{pmatrix}\varphi_0^T(x_i)\\ \vdots\\ \varphi_{K-1}^T(x_i)\end{pmatrix}$ be the feature matrix of the model, $L(\theta)=\frac{1}{2\sigma^2}(y-\Phi_K\theta)^T(y-\Phi_K\theta)$, $\theta^*=(\Phi_K^T\Phi_K)^{-1}\Phi_K^Ty$, $\nabla L(\theta):=\frac{1}{\sigma^2}\Phi_K^T(\Phi_K\theta-y)$

2) $\Phi_K(x)=\text{Vandermonde}(x,K)^T$

ROOT MEAN SQUARE ERROR (RMSE): measure of error for the model, $RMSE:=\sqrt{\frac{1}{N}\|y-\Phi\theta\|^2}=\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n-\Phi^T(x_n)\theta)^2}$

MAXIMUM A POSTERIORI ESTIMATION (MAP): reduces MLE's overfitting, assume $\exists$ a prior probability $P(\theta)$ on the parameters then $P(\theta|x,y)=\frac{P(y|x\theta)\cdot P(\theta)}{P(y|x)}$, $\theta^*=\operatorname{argmax}P(\theta)$

N.B.: 1) usually $P(\theta)\sim N(0,\sigma^2)$ to limit the parameters' magnitude

2) the evidence $P(y|x)$ is constant w.r.t. θ, negligible

3) assuming $P(\theta)\sim N(0,\sigma^2 I)$, $L(\theta)=-\log P(\theta|x,y)=-\log P(y|x,\theta)-\log P(\theta)=\frac{1}{2\sigma^2}\|x\theta-y\|^2+\frac{1}{2\sigma^2}\|\theta\|^2=\frac{1}{2}\|x\theta-y\|^2+\frac{\lambda}{2}\|\theta\|^2$ where $\lambda=\frac{\sigma^2}{\sigma^2}$ (=regularization parameter)

4) $L(\theta)$ convex, $\nabla L(\theta)=0\iff(x^Tx\theta-x^Ty)+\lambda\theta=0\iff(x^Tx+\lambda I)\theta=x^Ty$