Linear Algebra

Vector Properties

Linear independence — Linear combination $Au=u_1a_1+...+u_na_n=\sum_{i=1}^n u_ia_i$, where A is a vector matrix and u provides scaling values, is linearly independent if any of the following holds: • If equation system $Au=0$ then $u=0$ • A is full rank

Unique representation theorem: Any vector v that can be represented by a set of linearly independent vectors $a_1,...,a_n$ has a unique representation $v=\sum_{i=1}^n u_ia_i$ in terms of these vectors

Unit vector — $u=\frac{u}{\|u\|}$, therefore $\|u\|^2=1$

Inner product — $u\cdot v=u^\top v=\sum_{i=1}^n u_iv_i=cos(\varphi)\|u\|\|v\|$ resp. $\langle u,v\rangle_W=u^\top Wv=\sum_i u_iw_iv_i$ where W is either a diagonal matrix with $w_i>0$ or a pd matrix. Properties: • $u\cdot v=v\cdot u$ • $(u+v)\cdot w=u\cdot w+v\cdot w$ • $(\alpha u)\cdot v=\alpha(u\cdot v)$ • Positive definite: $u\cdot u\ge0$ • $u\cdot u=0\Leftrightarrow u=0_V$

ℓ_2 norm — $\|u\|=\sqrt{u\cdot u}$ resp. $\|u\|_W^2=u^\top Wu=\sum_i u_i^2w_i$ where W is a diagonal matrix with $w_i>0$. Properties: • $\|\alpha u\|=|\alpha|\|u\|$ • Positive definite (see inner product) • Equals 0 for zero vector (see inner product)

ℓ_1 norm — $|u|=\sum_i|u_i|$

Distance and angle between two vectors — • Distance: $d=\|u-v\|=\sqrt{(u_1-v_1)^2+...+(u_n-v_n)^2}$ • Angle: $cos(\varphi)=\frac{u\cdot v}{\|u\|\|v\|}$

Cauchy Schwarz ineq. — $|u\cdot v|\le\|u\|\|v\|$ with eq. iff $\varphi=0$ i.e. $u=\alpha v$ or if u or $v=0_V$

Triangle ineq. — $\|u+v\|\le\|u\|+\|v\|$ resp. $\|u-v\|\le\|u\|+\|v\|$ with eq. iff $\varphi=0$ i.e. $u=\alpha v$ or if u or $v=0_V$

Other inequalities — • $\|n^k\|\le\|n\|^k$ • $|\sum_i n_i|\le\sum_i|n_i|$

Orthogonal vectors — Properties: • $u\cdot v=0$ resp. $u^\top Wv=0$ • $\|u+v\|^2=\|u\|^2+\|v\|^2$ • *Pythagorean theorem*: $\|u-v\|^2=\|u\|^2+\|v\|^2$ • Orthogonal vectors are linearly independent, but not vice versa

Orthonormal vectors — Vectors are orthonormal iff $\|u\|=\|v\|=1$ and $u\cdot v=0$

Projection — Projection of $v\in V$ onto $s\in S$ given by: $v_S=\frac{v\cdot s}{\|s\|^2}s$ and if s is a unit vector $=(v\cdot s)s$

Vector Spaces

Vector space V

Subspace S

Invariant subspace H — H is an invariant subspace of S spanned by S if $Sh\in H$ for all $h\in H$. Properties: • S has an eigenvector in H • If S is symmetric, $H^\perp$ is also an invariant subspace of S

Orthogonal complement $S^\perp$ — Subspace, composed of set of vectors that are orthogonal to S. Properties: • The intersection of S and $S^\perp$ is $\{0_V\}$ • $\dim(S)+\dim(S^\perp)=\dim(V)$

Span — • Span of $\{s_i\}_{i=1}^n$ is the set of all vectors that can be expressed as a linear combination of $\{s_i\}_{i=1}^n$: $\sum_{i=1}^n u_is_i$ • Span of matrix A is the span of its column vectors: $Au=u_1a_1+...+u_na_n=\sum_{i=1}^n u_ia_i$ • A span is a subspace

(Orthonormal) basis — Unique set of all (orthonormal) vectors $\{s_i\}_{i=1}^n$ that are linearly independent and span the whole of a subspace. • *Orthonormal representation theorem*: Any vector $x\in S$ can be expressed as linear combination of resp. projection to orthonormal basis: $x=\sum_i(x\cdot s_i)s_i$ • *Parseval's theorem*: Extension of orthonormal representation theorem: $x\cdot y=\sum_i(x\cdot s_i)(y\cdot s_i)$ resp. $\|x\|^2=\sum_i|(x\cdot s_i)|^2$ • *Gram Schmidt orthonormalization*: Procedure to generate orthonormal basis $\{s_i\}_{i=1}^n$ from linearly independent vectors $\{x^{(i)}\}_{i=1}^n$: 1) $s_1=\frac{x_1}{\|x_1\|}$ 2) $\tilde s_k=x_k-\sum_{i=1}^{k-1}(x_k\cdot s_i)s_i$ for $k>1$ 3) $s_k=\frac{\tilde s_k}{\|\tilde s_k\|}$

Dimension d — • A vector space is *finite-dimensional* if $V=span(S)$ • $V\in\mathbb{R}^d$ has d basis vectors and each vector has d elements

Orthogonal vectors in spaces — • Let S be spanned by orthonormal $s_1,s_2,...$ and $v\in V$ • *Orthogonal decomposition theorem*: $v=v_S+v_{S^\perp}$ where $v\in V$, $v_S\in S$ and $v_{S^\perp}\in S^\perp$ • *Orthogonality principle* ∘ v_S is the projection of $v\in V$ to S iff $(v-v_S)\cdot s_i=0$ ∘ This can be rewritten to linear equation system $v\cdot s_i=v_S\cdot s_i=\sum_k\alpha_k(s_k\cdot s_i)$ since v_S can be expressed as linear combination of resp. projection to orthonormal basis $v_S=\sum_k\alpha_ks_k$ • $v_{S^\perp}=v-v_S=v-\sum_k(v\cdot s_k)s_k$ • *Approximation in a subspace theorem*: ∘ Unique best representation of v in S is given by projection of v to S: $\|v-s'\|\ge\|v-v_S\|$ for some arbitrary $s'\in S$ ∘ Any subset U of S is closest to v iff it is closest to v_S ■ We have: ∘ $\|v_S+v\|^2=\|v_S\|^2+\|v\|^2+2v_S\cdot v=\|v_S\|^2+\|v\|^2+2v_S\cdot(v_S+v_{S^\perp})=3\|v_S\|^2+\|v\|^2$ ∘ $\|v_S-v\|^2=\|v_S\|^2+\|v\|^2-2v_S\cdot v=\|v_S\|^2+\|v\|^2-2v_S\cdot(v_S+v_{S^\perp})=\|v\|^2-\|v_S\|^2$

Linear Equations

Let $Xb=y$ where $X\in\mathbb{R}^{n\times m}$, $b\in\mathbb{R}^m$, $y\in\mathbb{R}^n$ and b is unknown • Number of distinct equations = Number of linearly independent rows in $[X|b]$ = rank$([X|b])\le\min(n,m+1)$ • Number of LHS solutions should = Number of RHS solutions = rank$(X)\le\min(n,m)$ Solutions: • If rank$(X)<$rank$([X|b])$, system is inconsistent (no sol.) • If rank(X)=rank$([X|b])<m$, system is singular (infinitely many solutions) and underdetermined because we have fewer distinct equations than unknowns • If rank(X)=rank$([X|b])=m=n$, system is non-singular (exactly one sol.) and exactly determined • If rank(X)=rank$([X|b])=m<n$, system is non-singular and overdetermined

General Matrix Properties

Matrices — • $A\in\mathbb{R}^{n\times m}$ with elements A_{ij}, rows $i=1,...,n$, columns $j=1,...,m$ • Transpose $A^\top$ • Identity matrix I with 1 on diagonal, 0 elsewhere • Scalar matrix K with k on diagonal, 0 elsewhere

Operations — Matrix multiplication: • $A^{n\times p}B^{p\times m}=C^{n\times m}$ ∘ $r_V\times c_V=s$ ∘ $c_V\times r_V=M$ ∘ $M\times c_V=c_V$ ∘ $r_V\times M=r_V$ ∘ $M\times M=M$ • Element in C is sum-product of row in A and column in B: $C_{ij}=A^{(i)}\cdot B^{(j)}$ • Column vector in C is a linear combination of the columns in A: $C^{(j)}=AB^{(j)}=\sum_pA^{(j=p)}b_p^{(j)}$ • Row vector in C is a linear combination of the rows in B: $C^{(i)}=A^{(i)}B=\sum_pa_p^{(i)}B^{(i=p)}$

• $C=A[B^{(j=1)}|...|B^{(j=m)}]$ • $C=[A^{(i=1)}|...|A^{(i=n)}]^\top B=[A^{(i=1)}B|...|A^{(i=n)}B]^\top$

• $Ae_k=A^{(j=k)}$ and $e_k^\top A=A^{(i=k)}$ where e_k = 1 on k^{th} element and 0 everywhere else • $u\cdot v=u^\top v=\sum_iu_iv_i=c$

• $uv^\top=C$ with $u_iv_j=C_{ij}$ • $Au=\sum_{j=i}A^{(j)}u_i=c$ with $A^{(i)}u=c_i$ • $u^\top A=\sum_{j=i}A^{(j)}u_j=c^\top$ with $u\cdot A^{(j)}=c_j$

• $AB=\sum_{j=i}A^{(i)}B^{(j)}=C$ with $A^{(i)}B^{(j)}=C_{ij}$

• $u^\top Au=\sum_i\sum_jx_iA_{ij}x_j$, which we can specify: ∘ If A is symmetric: $=\sum_{i<j}x_i(A_{ij}+A_{ji})x_j+\sum_iA_{ii}x_i^2$, since $A_{ij}=A_{ji}$ and so we can only sum over $i\le j$ ∘ If A is diagonal: $=\sum_ix_iA_{ii}x_i$ since all off-diagonal terms are 0

• $(Au)_i=u\cdot A^{(i)}=\sum_jA_{ij}u_j$

• $(AB)_{ij}=A^{(i)}B^{(j)}=\sum_kA_{ik}B_{kj}$ • Moving between instance-level → data-level: ∘ $x^{(i)}y=a\to X^\top y=a$ where X consists of rows $x^{(i)}$ ∘ $x^{(i)}x^{(i)\top}=A\to X^\top X=A$ where X consists of rows $x^{(i)}$ ∘ $x^{(i)}\cdot\beta=y_i\to X\beta=y$ where X consists of rows $x^{(i)}$

Properties — • $(A+B)^\top=A^\top+B^\top$ • $(\alpha A)^\top=\alpha A^\top$ • $(AB)^\top=B^\top A^\top$ • $(A^\top)^k=(A^k)^\top$ • $A(B+C)=AB+AC$ • $(A+B)C=AC+BC$ • $A+B=B+A$ • $\alpha(A+B)=\alpha A+\alpha B$ • $(\alpha+\beta)A=\alpha A+\beta A$ • $(\alpha\beta)A=\alpha(\beta A)$ • $(A+B)x=Ax+Bx=Cx$ • $(AB)x=A(Bx)=Cx$ • $A=0.5(A+A^\top)+0.5(A-A^\top)=B+C$ where B is symmetric, but not C • $A=AI=IA$ • $Ak=AK=KA$ • rank(AB)=min(rank(A),rank(B)) • $A^\top A$ satisfies: ∘ Symmetric ∘ Psd ∘ Has rank m iff it is pd ∘ Invertible iff it has rank m and it is pd ∘ rank$(A^\top A)$=rank$(A^\top)$ ∘ rank$(A^\top A)$=rank$([A^\top A|A^\top x])$ ∘ $A^\top A=0$ implies $A=0$. Proof: $x^\top A^\top Ax=\|Ax\|^2=Ax=0$ for $\forall x$ implies $A=0$

Matrix terminology — • *Kernel* null(X) contains set of vectors b such that linear map $Xb=0$ • *Nullity* = dim(null(X)) • *Image* range(X) contains set of vectors b generated by linear map Xb resp. is space spanned by columns of X • *Row space* is space spanned by rows of X • *Column rank* = dim(colspace(X)) = number of linearly independent columns (not scalar multiples of eachother) • *Row rank* = dim(rowspace(X)) = number of linearly independent rows (not scalar multiples of eachother) • *Rank* = column rank = row rank = dim(range(X)) = dim(range$(X^\top)$) $\le min(n,m)$ • For invertible B, colspace(XB)=colspace(X) and rowspace(XB)=rowspace(X) • *Rank nullity theorem*: Rank(X)+nullity$(X)=m$

Matrices as linear maps — X maps b from $\mathbb{R}^m$ to $\mathbb{R}^n$: $Xb=y$ with $X\in\mathbb{R}^{n\times m}$, $b\in\mathbb{R}^m$, $y\in\mathbb{R}^n$ • *Injective*: $Xb=y$ has at most one sol., happens iff columns of X are linearly independent (rank$(X)=m\le n$) • *Surjective*: $Xb=y$ has at least one sol., happens iff rows of X are linearly independent (rank$(X)=n\le m$) • *Bijective*: Mapping is both injective and surjective, i.e. $m=n$

Projection matrices — Generally: • Projection matrix satisfies $P=P^2$ • Proof: ∘ Let S be spanned by $\{y_i\}_{i=1}^n$, which are column vectors of the matrix $A\in\mathbb{R}^{m\times n}$ ∘ Then, Ac are linear combinations of $\{y_i\}_{i=1}^n$ ∘ A vector $(x-Ac)$ is orthogonal to the columnspace of A, if: columnspace$(A)\cdot(x-Ac)=A^\top(x-Ac)=A^\top x-A^\top Ac=0$ ∘ Then, $c=(A^\top A)^{-1}A^\top x$ ∘ With this definition of c, we have $A(A^\top A)^{-1}A^\top$ as the projection matrix P ∘ $P^2=(A(A^\top A)^{-1}A^\top)(A(A^\top A)^{-1}A^\top)=A(A^\top A)^{-1}A^\top=P$ • For projection onto line: $P=\frac{bb^\top}{b^\top b}$ resp. $P=\frac{Wbb^\top W}{b^\top Wb}$ for weighted inner product • For projection onto subspace: $P=B(B^\top B)^{-1}B^\top$ resp. $P=WB(B^\top WB)^{-1}B^\top W$ for weighted inner product Via orthonormal basis: Let S be spanned by orthonormal $\{b_i\}_{i=1}^n$, which are column vectors of the matrix $B\in\mathbb{R}^{m\times n}$ • Projection of x onto S is given by: $u=\sum_i(x\cdot b_i)b_i=\sum_ib_ib_i^\top x=BB^\top x=Cx$ • Projection of x onto $S^\perp$ is given by: $x-u=x-Cx$ Via SVD: Let S be spanned by $\{y_i\}_{i=1}^n$, which are column vectors of the matrix $A\in\mathbb{R}^{m\times n}$ • Projection of x onto S is given by: $s=AA^\#x$ since $AA^\#$ is a projection matrix due to $AA^\#=(AA^\#)^2$ • $s=U_+U_+^\top x$ • *SVD Projection Energy*: $\sum_{l=1}^m|u_k\cdot y_l|^2=\sigma_k^2$ Proof: $\sum_{l=1}^m|u_k\cdot y_l|=\|u_k^\top A\|=u_k^\top AA^\top u_k=u_k^\top USV^\top VS^\top U^\top u_k=e_k^\top SS^\top e_k=\sigma_k^2$

Square Matrix Properties

Square matrix terminology — • *Diagonal matrix*: ∘ Def: Has $\{d_i\}_{i=1}^n$ on diagonal and 0 everywhere else ∘ For diagonal matrices: $DD^\top=D^\top D$ • *Inverse matrix*: ∘ Def: $A^{-1}A=I$ ∘ Is unique ∘ For diagonal matrices: A^{-1} can be calculated by inverting all diagonal elements • *Symmetric (Hermitian) matrix*: ∘ $A^\top=A$ ∘ Properties: ■ $(x+A^{-1}b)^\top A(x+A^{-1}b)-b^\top A^{-1}b=x^\top Ax+2x^\top b$ ■ If A and B are symmetric, $A+B$ is also symmetric • *Orthogonal (unitary) matrix*: ∘ Def: $A^\top=A^{-1}$ ∘ $AA^\top=A^\top A=I$ ∘ Rows and columns are orthonormal ∘ $\|Ax\|=\|x\|$ ∘ $(Ax)\cdot(Ay)=x\cdot y$ • *Involution matrix*: $A^{-1}=A$ • *Determinant*: ∘ Function which maps A to a scalar ∘ Properties: ■ det(I)=1 ■ det(AB)=det(A)det(B) ■ det$(A^\top)$=det(A) ■ det(A^{-1})=(det$(A))^{-1}$ ■ det$(\alpha A)=\alpha^2$det(A) ■ Determinant of diagonal matrix is product of diagonal elements ■ det$\begin{bmatrix}a&b\\c&d\end{bmatrix}=ad-bc$

Invertible matrix theorem — Following statements are equivalent for square matrix $A\in\mathbb{R}^{n\times n}$: • A is invertible • Only sol. to $Ax=0$ is $x=0_V$ Proof: ∘ $A^{-1}Ax=0\Rightarrow Ix=0\Rightarrow x=0_V$ • A is non-singular • Columns (and rows) of A are linearly independent • rank$(A)=n$ • det$(A)\ne0$ • Singular values of A are strictly positive Inversely, if A is not invertible, the columns and rows are not linearly independent, etc.

Matrix inversion lemma — • Let $B\in\mathbb{R}^{n\times n}$, $D\in\mathbb{R}^{m\times m}$, $C\in\mathbb{R}^{n\times m}$. Then, $A=B+CD^{-1}C^\top$ is invertible: $A^{-1}=B-BC(D+C^\top BC)^{-1}C^\top B$ • Let $v\in\mathbb{R}^n$. Then, $(\alpha I+vv^\top)^{-1}v=(\alpha+\|v\|^2)^{-1}v$ • $v^\top(\alpha I+vv^\top)^{-1}=v^\top(\alpha+\|v\|^2)^{-1}$

Eigenvectors and eigenvalues — • q is an eigenvector of A associated with an eigenvalue λ if it remains on the same line after transformation by a linear map: $Aq=\lambda q$ • Let $A\in\mathbb{R}^{n\times n}$. A can have between $1-n$ eigenvalues, each with multiple eigenvectors. Eigenvectors for distinct eigenvalues are linearly independent • *Spectral radius*: $\rho(A)$ is the largest eigenvalue of A • If there exists a non-trivial sol. for q, $(A-\lambda I)$ is not invertible and characteristic polynomial det$(A-\lambda I)=0$ • *Eigendecomposition resp. diagonalization*: $A=Q\Lambda Q^{-1}$ where Q is a matrix with the eigenvectors as columns and Λ is a diagonal matrix with the eigenvalues on the diagonal • det(A)=det$(Q\Lambda Q^{-1})=\prod_{i=1}^n\lambda_i$ • *Symmetric eigendecomposition resp. unitary diagonalization*: For symmetric A: $A=Q\Lambda Q^\top$ where Q is an orthogonal matrix with the eigenvectors as columns and Λ is a diagonal matrix with the eigenvalues on the diagonal • *Spectral theorem*: Square matrix A is symmetrically diagonalizable, iff $AA^\top=A^\top A$ • *Spectral theorem for symmetric matrices*: Every symmetric matrix A is symmetrically diagonalizable (due to Spectral theorem) and all its eigenvalues are real

Positive definite (pd) and positive semi-definite matrices (psd) — • $A>0$ iff $x^\top Ax>0$ • $A\ge0$ iff $x^\top Ax\ge0$ Properties: • If A is p(s)d, A^{-1} is also p(s)d • If A and B are p(s)d, $A+B$ is also p(s)d • If det$(A)=\prod_{i=1}^n\lambda_i>(\ge)$ 0 resp. $\{\lambda_i\}_{i=1}^n>(\ge)$ 0 for pd (psd) Pd properties: • I is pd • If A is pd, A^{-1} is pd • *Cholesky decomposition*: If A is pd, $A=BB^\top$ • If A and B are pd, $(AB)^{-1}=B^{-1}A^{-1}$ Psd properties: • If A is psd, $BAB^\top$ is psd

Singular Value Decomposition (SVD)

SVD — • For $A\in\mathbb{R}^{n\times m}$, orthogonal rotation matrix $U\in\mathbb{R}^{n\times n}$, diagonal scaling and projection matrix $S\in\mathbb{R}^{n\times m}$, and orthogonal rotation matrix $V\in\mathbb{R}^{m\times m}$: $A=USV^\top$ • For symmetric $A\in\mathbb{R}^{n\times n}$: $A=USU^\top$ • In S: ∘ Diagonal elements $\sigma_1,...$ are the *singular values* of A ∘ If $\sigma_1\ge\sigma_2...\ge0$, S is unique ∘ *Spectral norm* = $\sigma_{max}=\|A\|_{\text{operator}}=sup_{x\ne0}\frac{\|Ax\|_2}{\|x\|_2}$ Proof: ■ $A=U\Sigma V^T$ ■ For any vector $x\in\mathbb{R}^N$, we have: $\|Ax\|_2=\|U\Sigma V^Tx\|_2$ ■ Since U is orthogonal, we can write: $\|Ax\|_2=\|\Sigma V^Tx\|_2$ ■ Let $y=V^Tx$. Substituting, we get: $\|Ax\|_2=\|\Sigma y\|_2$ ■ The diagonal matrix Σ scales the components of y by the singular values σ_i: $\|\Sigma y\|_2=\sqrt{\sum_{i=1}^r(\sigma_iy_i)^2}$ ■ The supremum of $\|\Sigma y\|_2$ occurs when all the weight is on the largest singular value σ_1 ■ Then, we see that: $\|A\|_2=\sup_{y\ne0}\frac{\|\Sigma y\|_2}{\|y\|_2}$ is achieved when y is aligned with the singular vector corresponding to σ_1, giving: $\|A\|_2=\sigma_1=\sigma_{\max}(A)$ ∘ Largest singular value σ_{max} is always greater than largest eigenvalue $\rho(A)$ ∘ *Condition number* $=\sigma_{max}/\sigma_{min}$ ∘ For square A: Iff $\sigma_1,\sigma_2,...>0$, A is invertible ∘ SVD is closely related to spectral theorem: ∘ According to spectral theorem, every matrix A is symmetrically diagonalizable (i.e. $A=Q\Lambda Q^\top$), iff $AA^\top=A^\top A$ ∘ If we apply SVD to $AA^\top$ resp. $A^\top A$: ■ $AA^\top=USV^\top VS^\top U^\top=U(SS^\top)U^\top$ since V is orthogonal and $V^\top V=I$ ■ $A^\top A=VS^\top U^\top USV^\top=V(S^\top S)V^\top$ since U is orthogonal and $U^\top U=I$ ∘ $SS^\top$ and $S^\top S$ are diagonal matrices with elements $\sigma_1^2,\sigma_2^2,...$ ∘ Given symmetric diagonalization for any matrix, we see that ■ S with σ_i contains square root of eigenvalues of $AA^\top$ resp. $A^\top A$ ■ U contains eigenvectors of $AA^\top$ as columns resp. V contains eigenvectors of $A^\top A$ as columns ∘ According to spectral theorem, symmetric matrix A is symmetrically diagonalizable (i.e. $A=Q\Lambda Q^\top$) ∘ If we apply SVD to symmetric matrix A, we see that ■ S contains absolute value of eigenvalues of A ■ U contains eigenvectors of A as columns

Pseudo Inverse — • Pseudo Inverse satisfies certain conditions that make it behave like an inverse for matrices that might not be invertible in the usual sense • $A^\#=VS^\#U^\top$ • $S^\#=\begin{bmatrix}S_+^{-1}&\cdots&0\\0&\cdots&0\end{bmatrix}$ is obtained from S by first transposing it and then taking the inverse of non-zero singular values • $A^\#$ is unique • If $rank(A)$ = number of rows in A then: ∘ $S^\top$ and $S^\#$ have full column rank and $S^\#=S^\top(SS^\top)^{-1}$ ∘ $AA^\#=I$ ∘ $A^\#=A^\top(AA^\top)^{-1}$ ∘ Pseudo inverse provides minimum norm sol., when system $y=Ax$ is underdetermined: $x=A^\top(AA^\top)^{-1}y$ • If $rank(A)$ = number of columns in A then: ∘ $S^\top$ and $S^\#$ have full row rank and $S^\#=(S^\top S)^{-1}S^\top$ ∘ $A^\#A=I$ ∘ $A^\#=(A^\top A)^{-1}A^\top$ ∘ Pseudo inverse provides least squares sol., when system $y=Ax$ is overdetermined: $x=(A^\top A)^{-1}A^\top y$

Properties — • $AA^\#A=A$ • $A^\#AA^\#=A^\#$ • $(A^\top)^\#=(A^\#)^\top$ • $(AA^\top)^\#=(A^\#)^\top A^\#$ • $A^\#x=0\Leftrightarrow A^\top x=0$ • Properties can be proven by replacing A by its SVD and $A^\#$ by its definition • Column space of $A^\#$ equals column space of $A^\top$ • $AA^\#=US^\#U^\top=U_+U_+^\top$ • Property can be proven by replacing A and $A^\#$ by their SVD • $SS^\#=I_+$

Hilbert Space S

Equivalence modulo norm zero: • Challenge: In some cases, $v\cdot v=\Leftrightarrow v=0_V$ does not hold • Issue is resolved by defining equivalence classes of vectors: $[v]=\{v'\in V:\|v-v'\|=0\}$ • Implications: Modified meaning of eq.: ∘ For functions: $f=g$ means $\int_T|f(t)-g(t)|^2dt=0$ ∘ For random variables: $X=Y$ means $\mathbb{E}[|X-Y|^2]=0$ Existence (convergence) of the inner product: • Challenge: In some cases, inner product does not exist for all $v,w\in V$ • Issue is resolved by restricting attention to subsets of V where the norm is finite: $V_{fn}=\{v\in\|v\|<\infty\}$ Hilbert spaces: • Vector space with an inner product that satisfies the additional condition of *completeness*: ∘ Every Cauchy sequence in V converges to an element in V resp. limit vectors, that Cauchy sequences tend towards, are also elements of V ∘ *Cauchy sequence*: Sequence of points that get closer and closer • If we make modifications for above challenges (equivalence modulo norm zero, existence of inner product), vector spaces can be transformed to Hilbert spaces

Other

Common exp and log rules — • $a^m\cdot a^n=a^{m+n}$ • $\frac{a^m}{a^n}=a^{m-n}$ • $(ab)^n=a^nb^n$ • $(\frac{a}{b})^n=\frac{a^n}{b^n}$ • $a^{-n}=\frac{1}{a^n}$ • $a^0=1$ • $a^1=a$ • log(xy)=log$x$+log$y$ • log$(\frac{x}{y})$=log$x$−log$y$ • log$(x^n)=n$ log$x$ • log 1=0 • log$(x<1)<0$ • log$(x>1)>0$ • $e^{log(x)}=log(e^x)=x$

Geometric series — • Finite: $S_n=\sum_{i=1}^na_ir^{i-1}=a_1(\frac{1-r^n}{1-r})$ • Infinite: $S=\sum_{i=0}^\infty a_ir^i=\frac{a_1}{1-r}$ for $r<1$

Polynomials — $\sum_ia_ix^i$

Calculus

Derivatives

Rules — • Product rule: $\frac{\partial f\times g}{\partial x}=f\times\frac{\partial g}{\partial x}+g\times\frac{\partial f}{\partial x}$ • Quotient rule: $\frac{\partial\frac{f}{g}}{\partial x}=\frac{g\frac{\partial f}{\partial x}-f\frac{\partial g}{\partial x}}{g^2}$ • Chain rule: $\frac{\partial f(g)}{\partial x}=\frac{\partial f}{\partial g}\times\frac{g}{\partial x}$ • Multivariate chain rule: For $f(z,v),z=g(x),v=h(x)$: $\frac{\partial f}{\partial x}=\frac{\partial f}{\partial z}\frac{\partial z}{\partial x}$

Common derivatives — • $\frac{\partial x^n}{\partial x}=nx^{n-1}$ • $\frac{\partial e^{kx}}{\partial x}=k\times e^{kx}$ • $\frac{\partial log(x)}{\partial x}=\frac{1}{x}$ • $\frac{\partial\sqrt{x}}{\partial x}=\frac{1}{2\sqrt{x}}$ • $\frac{\partial\sin(x)}{\partial x}=\cos(x)$ • $\frac{\partial\cos(x)}{\partial x}=-\sin(x)$

Partial and directional derivative — • For a funct. that depends on n variables $\{x_i\}_{i=1}^n$, partial derivative is slope of tangent line along direction of one specific variable x_i • Directional derivative is slope of tangent line along direction of selected unit vector u

Gradient — • Given scalar-valued funct. $f:\mathbb{R}^n\to\mathbb{R}$, returns vector containing first-order partial derivatives: $\nabla_xf:[\frac{\partial f}{\partial x_1}\cdots\frac{\partial f}{\partial x_n}]^\top$ • Gradient points in direction of greatest upward slope of f • Magnitude of gradient equals rate of change when moving into direction of greatest upward slope

Hessian — • Given scalar-valued funct. $f:\mathbb{R}^n\to\mathbb{R}$, returns matrix containing second-order partial derivatives: $\mathcal{H}=\nabla_x^2f:\begin{bmatrix}\frac{\partial^2f}{\partial x_1^2}&\cdots&\frac{\partial^2f}{\partial x_1\partial x_n}\\\vdots&\ddots&\vdots\\\frac{\partial^2f}{\partial x_n\partial x_1}&\cdots&\frac{\partial^2f}{\partial x_n^2}\end{bmatrix}\in\mathbb{R}^{n\times n}$ • $\mathcal{H}$ is symmetric

Jacobian — • Given vector-valued funct. $f:\mathbb{R}^n\to\mathbb{R}^m$ with $f=[f_1(x),...,f_m(x)]^\top$, returns matrix containing first-order partial derivatives: $\nabla_xf:\begin{bmatrix}\frac{\partial f_1}{\partial x_1}&\cdots&\frac{\partial f_1}{\partial x_n}\\\vdots&\ddots&\vdots\\\frac{\partial f_m}{\partial x_1}&\cdots&\frac{\partial f_m}{\partial x_n}\end{bmatrix}\in\mathbb{R}^{m\times n}$

Matrix calculus rules — • $\frac{\partial x^\top x}{\partial x}=2x$ • $\frac{\partial a^\top x}{\partial x}=a$ • $\frac{\partial Ax}{\partial x}=A^\top$ • $\frac{\partial x^\top Ax}{\partial x}=(A+A^\top)x$ • $\frac{\partial a^\top Ab}{\partial A}=ab^\top$ • For symmetric A: $\frac{\partial x^\top Ax}{\partial x}=2Ax$ • For square A: ∘ $\frac{\partial a^\top A^{-1}b}{\partial A}=-(A^\top)^{-1}ac^\top(A^\top)^{-1}$ ∘ $\frac{\partial log(|A|)}{\partial A}=(A^\top)^{-1}$

Extrema

Conditions for local min. and max. — • Point is a stationary point, i.e. first-order derivative =0 • If Hessian is pd, it's a local min., if Hessian is nd, it's a local max., if Hessian is indefinite, it's a saddle point • Local min. and max. are the unique global min. and max. in strictly convex functions resp. one of possibly infinitely many global min. and max. in convex functions

Convexity — • For a convex funct.: ∘ $f(\lambda x+(1-\lambda)y)\le\lambda f(x)+(1-\lambda)f(y)$ with $\lambda\in[0,1]$ ∘ Hessian of stationary point(s) is psd ∘ Global min. exists, but may not be unique • For a strictly convex funct.: ∘ $f(\lambda x+(1-\lambda)y)<\lambda f(x)+(1-\lambda)f(y)$ with $\lambda\in[0,1]$ ∘ Hessian of stationary point is pd ∘ Unique global min. exists • Sum of convex functions $f_2(x)+f_1(x)$ is also convex, sum of convex and strictly convex funct. is strictly convex • Composition of convex functions $f_2(f_1(x))$, where outer funct. $f_2(x)$ is increasing, is also convex • Scalar multiple of convex funct. $\lambda f(x)$, where $\lambda\ge0$, is also convex • Any norm is convex • From this, we can derive that reg. risk is convex

Nature of optimum — What does Hessian and funct. look like? • If Hessian is pd and loss funct. is strictly convex, stationary point is a global min., and there is a unique sol. • If Hessian is psd and loss funct. is convex, stationary point is a global min., and there may be a geometrically unique or infinitely many solutions • If Hessian is p(s)d but loss funct. is not convex, stationary point may be a local min. and there may be a geometrically unique or infinitely many solutions

Opt. approach — Is funct. differentiable, continuous, and are relevant terms invertible? • If yes, analytically solvable • If no, numerically solvable (e.g. via gradient descent)

Constrained opt. — • *Lagr. funct.*: $\mathcal{L}(x,\lambda)=f(x)+\lambda g(x)$, where $g(x)$ is an $(m-1)$ dimensional constr. surface and λ is the Lagrange multiplier • $\nabla_x\mathcal{L}=\nabla_xf(x)+\lambda\nabla_xg(x)$ • $\nabla_\lambda\mathcal{L}=g(x)$ • Solution is feasible if it fulfills constr. and optimal, if no other feasible sol. produces a lower error • Min. over Lagr. corresponds to min. log-loss: ∘ $\hat x$=arg max$_xp(D|x)\rho(x)$ =arg min$_x(-log\,p(D|x)+k(x))$ ∘ where $k(x)=-log\rho(x)$ For eq. constr.: Min. $f(x)$ subject to $g(x)=0$ • Gradient of $f(x)$ must be orthogonal to constr. surface, otherwise (if it points into any direction along the constr. surface) $f(x)$ could still decrease for movements along the constr. surface • On the constr. surface, $g(x)$ is a constant, so moving along any direction on the constr. surface has a directional derivative of 0. Since the gradient of $g(x)$ points into the direction of steepest ascent, it must be orthogonal to the constr. surface, otherwise (if it points into any direction along the constr. surface) $g(x)$ would not be constant on the constr. surface • Then, gradients are parallel at optimum: $\nabla_xf(x^*)=\lambda^*\nabla_xg(x^*)$ • To find x^* and λ^*: ∘ $\nabla_x\mathcal{L}$=0, expresses parallelity condition at min. x^* ∘ $\nabla_\lambda\mathcal{L}$=0, expresses constr. ∘ This is an unconstr. opt. prob. • Optimum x^* and λ^* represents a saddle point of $\mathcal{L}$ For ineq. constr.: Min. $f(x)$ subject to $g(x)\le0$ • If x^* lies in $g(x)<0$, constr. is inactive • Otherwise, if x^* lies in $g(x)=0$, constr. is active: ∘ Gradient of $f(x)$ must point towards $g(x)<0$ region, otherwise (if it would point away from $g(x)<0$ region) the optimum would lie in this region ∘ Then, gradients are anti-parallel at optimum: $\nabla_xf(x^*)=-\lambda\nabla_xg(x^*)$ • To find x^* and λ^*: ∘ $\nabla_x\mathcal{L}$=0 subject to *Karush Kuhn Tucker conditions*: ■ $g(x)\le0$ ■ $\lambda\ge0$ ■ *Complementary slackness condition*: $\lambda g(x)=0$, with $\lambda=0,g(x)<0$ for inactive constr. and $\lambda>0,g(x)=0$ for active constr. ∘ $\nabla_\lambda\mathcal{L}$=0 given complementary slackness condition ∘ This is not an unconstr. opt. prob., but can be solved via duality • Optimum x^* and λ^* represents a saddle point of $\mathcal{L}$ For multiple constr.: Min. $f(x)$ subject to m ineq. constr. $\{g^{(i)}(x)\le0\}_{i=1}^m$ and p eq. constr. $\{h^{(j)}(x)=0\}_{j=1}^p$ • Then, Lagr. is given by: $\mathcal{L}(x,\lambda,\mu)=f(x)+\sum_{i=1}^m\mu^{(i)}g^{(i)}(x)+\sum_{j=1}^p\lambda^{(j)}h^{(j)}(x)$ • Then, general sol. x^*,λ^*,μ^* is given by: $\nabla_x\mathcal{L}$=0 subject to: ∘ $\{g^{(i)}(x)\le0\}_{i=1}^m$ and $\{h^{(j)}(x)=0\}_{j=1}^p$ ∘ $\{\mu^{(i)}\ge0\}_{i=1}^m$ ∘ $\{\mu^{(i)}g^{(i)}(x)=0\}_{i=1}^m$ *Primal prob.*: • min$_x[f(x)]$ s.t. constr. • Equivalent to min$_x$[max$_{\lambda,\mu}\mathcal{L}$]= min$_x[f(x)+max_{\lambda,\mu}[\sum_{i=1}^m\mu^ig^i(x)+\sum_{j=1}^p\lambda^jh^j(x)]]$ where term 2 is barrier function: ∘ =0 subject to constr. being met, given complementary slackness condition for ineq. constr. and $h^{(j)}(x)=0$ for eq. constr., which implies that dual prob. becomes min$_x(f(x))$ ∘ $=\infty$ otherwise, which implies that primal prob. cannot be solved *Dual prob.*: max$_{\lambda,\mu}$[min$_x\mathcal{L}$] *Weak duality*: • Always holds • Given mini-max theorem, dual sol. is lower bound of primal sol.: min$_x$[max$_{\lambda,\mu}\mathcal{L}$]=min$_x[f(x)]\ge$max$_{\lambda,\mu}$[min$_x\mathcal{L}$] • min$_x\mathcal{L}$ is an unconstr. opt. prob. • max$_{\lambda,\mu}$[min$_x\mathcal{L}$] is a concave max. prob. *Strong duality*: • Holds under *Slater's cond.* if there exists a strictly feasible point that fulfills all ineq. constr. $\{g^i(x)<0\}_{i=1}^m$ i.e. lin. sep. • Dual sol. equals primal sol.: min$_x$[max$_{\lambda,\mu}\mathcal{L}$]=min$_x[f(x)]$=max$_{\lambda,\mu}$[min$_x\mathcal{L}$] • Solve min$_x\mathcal{L}$ for primal variables x^* in terms of Lagr. multipliers λ,μ • Plug primal variables back into $\mathcal{L}$ and formulate additional constraints for optimum • Solve max$_{\lambda,\mu}\mathcal{L}$ for Lagr. multipliers λ^*,μ^* • Specify x^* based on λ^*,μ^*

Integrals

Indefinite integral — • $F(x)=\int f(x)dx$ *Definite integral* — • $F(b)-F(a)=\int_a^bf(x)dx$

Common integrals — • $f(x)=\frac{1}{a}x^n\to F(x)=\frac{1}{a}\frac{x^{n+1}}{n+1}$ for $n\ne1$ • $f(x)=\frac{1}{x}\to F(x)=log(x)$ • $f(x)=e^x\to F(x)=e^x$

Probability and Statistics

Terminology

Notation — • $A\cap B$ is the intersection of A and B, i.e. A and B • $A\cup B$ is the union of A and B, i.e. A or (inclusive) B

Kolmogorov axioms — *Probability space* defined by: • *Sample space*: All possible outcomes $\Omega=\{\omega_1,...,\omega_n\}$, e.g. for a dice toss $\{1,2,3,4,5,6\}$ • *Event space*: ∘ All possible results ∘ Corresponds to the *powerset* of the sample space: ■ Powerset includes the empty set, single-item sets, ..., full-item sets ■ E.g. powerset of $\{1,2,3,4,5,6\}$ is $\{\{\},\{1\},\{2\},...,\{1,2\},\{2,3\}...,\{1,2,3,4,5,6\}\}$ where e.g. event $\{1,2\}$ refers to the event of rolling a 1 or 2 ■ Powerset has size $2^{|\Omega|}$ ∘ An *event* is a subset of the sample space ∘ E.g. tossing an even number $\{2,4,6\}$ or $P(X\le r)$ • *Probability measure*: Function that assigns a probability to an event, e.g. p(tossing an even number)$=\frac{3}{6}=\frac{1}{2}$ Axioms: • Event space must be a *sigma algebra*: ∘ $\Omega\in$ event space ∘ If A is in event space, then its complement is also in event space with $P=1-a$ ∘ If $A_1...A_n$ are in event space with $P=a_1,...,a_n$, their union is also in event space with $P=a_1+...+a_n$ • Probability measure must satisfy: ∘ $0\le P(A)\le1$ ∘ $P(\Omega)=1$ ∘ If $A_1,A_2,...$ are in event space and do not intersect, then $P(A_1\cup A_2\cup...)=\sum_{n=1}^\infty P(A_n)$

Variables — • *Target space*: Numeric values that the random variable can take, e.g. for a dice toss $\{1,2,3,4,5,6\}$ • *Random variable*: ∘ Function that takes an element in sample space and returns a numeric value, e.g. $\mathcal{X}(3)=3$ ∘ *Discrete random variable*: Characterized by pmf, event given by $\{\omega\in\Omega|X(\omega)=s\}$ ∘ *Continuous random variable*: Characterized by pdf, event given by $\{\omega\in\Omega|X(\omega)\le r\}$ (or $>,=,\ge,<$, any unions and intersections) ∘ In general, if case is mixed (e.g. X is discrete, Y is continuous), then the joint probability is defined by the continuous terminology, the marginal probability defined by the terminology of the respective variable, and the conditional probability defined by the terminology of the respective dependent variable • *Independent random variables*: ∘ $P(A|B)=P(A)$ and $P(B|A)=P(B)$ ∘ $P(A\cap B)=P(A)P(B)$ resp. $F_{X_1,...,X_n}(r_1,...,r_n)=F_{X_1}(r_1),...,F_{X_n}(r_n)$ and $f_{X_1,...,X_n}(x_1,...,x_n)=f_{X_1}(x_1),...,f_{X_n}(x_n)$ ∘ Unnormalized correlation: $\mathbb{E}(\mathcal{X}\mathcal{Y})=\mathbb{E}(\mathcal{X})\mathbb{E}(\mathcal{Y})$ ∘ Covariance: Cov$(\mathcal{X},\mathcal{Y})$=0 ∘ Functions of independent random variables are also independent ∘ A subset of a larger set of independent random variables is also independent • *Conditionally independent random variables*: 2 random variables $\mathcal{X}$ and $\mathcal{Y}$ are conditionally independent, if there is a confounder $\mathcal{L}$ that causally affects both variables, but if we control for this confounder, the variables are not causally connected • *I.I.D. random variables*: Independent and from identical distribution • *Orthogonal random variables*: ∘ Unnormalized correlation: $\mathbb{E}(\mathcal{X}\mathcal{Y})$=0 ∘ Covariance not particularly defined ∘ Not necessarily independent • *Uncorrelated random variables*: ∘ Unnormalized correlation: $\mathbb{E}(\mathcal{X}\mathcal{Y})=\mathbb{E}(\mathcal{X})\mathbb{E}(\mathcal{Y})$ ∘ Covariance: Cov$(\mathcal{X},\mathcal{Y})$=0 ∘ Not necessarily independent

Events — • Complement: $P(A^C)=1-P(A)$ and $P(A\cup A^C)=P(A)+P(A^C)$ • *Disjoint / mutually exclusive vs. joint / mutually inclusive* • Subset $A\subset B$ with $P(A)<P(B)$

PMF, CDF, PDF — • *Cumulative density funct. (CDF)*: $F(r)=p(X\le r)$ • *Probability mass funct. (PMF)* for discrete random variables: $p(x)=p(X=x)$ • *Probability density funct. (PDF)* for continuous random variables: $f(x)$ • Properties of CDF and PDF: ∘ Derivative of CDF by x returns PDF: $f(x)=\frac{\partial F(x)}{\partial x}$ ∘ Integral of PDF by x returns CDF: $\int_{-\infty}^rf(x)dx=F(r)=p(X\le r)$ ∘ CDF is monotonically non-decreasing: If $s<r,F(s)<F(r)$ ∘ CDF is between 0 and 1: $lim_{r\to-\infty}F(r)=0$ and $lim_{r\to\infty}F(r)=1$ ∘ CDF is right-continuous: $lim_{s\to r^+}F(s)=F(r)$ ∘ For CDF: $lim_{s\to r^-}F(s)=F(x<r)=F(s)-F(x=r)$ ∘ $\int_a^bf(x)dx=F(b)-F(a)=p(a<X\le b)$ ∘ $\int_{-\infty}^\infty f(x)dx=1$

Probabilities — • Probability for single variable: *Marginal and total probability*: ∘ If X,Y are discrete: $p(x)=\sum_yp(x,y)=\sum_yp(x|y)p(y)$ ∘ If X,Y are continuous: $f(x)=\int_{-\infty}^\infty f(x,y)dy=\int_{-\infty}^\infty f(y)f(x|y)dy$ and $F(r)=\int_{-\infty}^r\int_{-\infty}^\infty f(x,y)dydx$ ∘ If X is discrete, Y is continuous: $p(x)=\int_{-\infty}^\infty p(x,y)dy=\int_{-\infty}^\infty p(x|y)f(y)dy$ ∘ If X is continuous, Y is discrete: $f(x)=\sum_yf(x,y)=\sum_yf(x|y)p(y)$ and $F(r)=\sum_yp(y)p(X\le r|y)=\sum_yp(y)F(r|y)$ • *Joint probability* $P(A,B)$: Probability for combination of variables ∘ If X is discrete: $p(x_1,...,x_n)$ ∘ If X is continuous: $f(x_1,...,x_n)=\frac{\partial^nF_{X_1,...,X_n}(x_1,...,x_n)}{\partial x_1,...,\partial x_n}$ and $F(r_1,...,r_n)=p(x_1\le r_1,...,x_n\le r_n)=\int_{-\infty}^{r_1}...\int_{-\infty}^{r_n}f_{X_1,...,X_n}(x_1,...,x_n)dx_n...dx_1$

• Conditional probability $P(A|B)=\frac{P(A\cap B)}{P(B)}$: Probability for variable, given other variable ○ If X,Y are discrete: $p(x|y)=\frac{p(x,y)}{p(y)}$ ○ If X,Y are continuous: $f(x|y)=\frac{f(x,y)}{f(y)}$ ○ If X is discrete, Y is continuous: $p(x|y)=\frac{f(x,y)}{f(y)}$ ○ If X is continuous, Y is discrete: $f(x|y)=\frac{f(x,y)}{p(y)}$ ○ Properties: ■ $P(A|B)=1-P(A^C|B)$ ■ $P(A_1|B)+P(A_2|B)+...=1$ ■ If conditioning on subset S: $p(x|S)=\begin{cases}p(x)/p(x\in S) & x\in S\\ 0 & x\notin S\end{cases}$ • Bayesian terminology: ○ Prior P(parameter) ○ Posterior P(parameter|data) ○ Likelihood P(data|parameter) ○ Evidence P(data) • Bayes theorem: Posterior $P(A|B)=\frac{\text{Likelihood } P(B|A)\times\text{Prior } P(A)}{\text{Evidence } P(B)}$ where $P(B)$ can be rewritten in marginalized form over A

Measures n^{th} *moment* — $\mathbb{E}(X^n)=\int_{-\infty}^{\infty}x^n f(x)dx$

Expected value — Generally: • If X is discrete: $\mathbb{E}(X)=\sum_X x\times p(x)$ • If X is continuous: $\mathbb{E}(X)=\int_{-\infty}^{\infty}x\times f(x)dx$ • If Y is discrete: $\mathbb{E}[X]=\sum_y \mathbb{E}[X|Y=y]p(y)$ • If Y is continuous: $\mathbb{E}[X]=\int_{-\infty}^{\infty}\mathbb{E}[X|Y=y]f(y)dy$ For functions: • $g(X)$ is a funct. • If X is discrete: $\mathbb{E}(g(X))=\sum_X g(x)\times p(x)$ • If X is continuous: $\mathbb{E}(g(X))=\int_{-\infty}^{\infty}g(x)\times f(x)dx$ For probabilities: • Count as functions • A is an event, X is a random variable • If X is discrete: $\mathbb{E}[p(X|A)]=\sum_X p(x|A)p(x)$ • If X is continuous: $\mathbb{E}[p(X|A)]=\int_{-\infty}^{\infty}f(x|A)f(x)dx$ • If X is discrete: $\mathbb{E}[p(A|X)]=\sum_X p(A|x)p(x)=\sum_X p(A,x)=p(A)$ • If X is continuous: $\mathbb{E}[p(A|X)]=\int_{-\infty}^{\infty}f(A|x)f(x)dx=p(A)$ For conditions: • A is an event, X is a random variable • If X is discrete: $\mathbb{E}(X|A)=\sum_X x\times p(x|A)$ • If X is continuous: $\mathbb{E}(X|A)=\int_{-\infty}^{\infty}x\times f(x|A)dx$ • $\mathbb{E}(X|A)=P(A|X)$ For vectors: • Expectation of a vector is the expectation of each of its elements • If X is discrete: $\mathbb{E}(\boldsymbol{x})=\sum_{x_1}\cdots\sum_{x_n}\boldsymbol{x}^\top p(x_1,...,x_n)=\boldsymbol{\mu}$ • If X is continuous: $\mathbb{E}(\boldsymbol{x})=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\boldsymbol{x}^\top f_{X_1,...,X_n}(x_1,...,x_n)dx_1...dx_n=\boldsymbol{\mu}$

Properties: ○ $\mathbb{E}(\alpha)=\alpha$ • $\mathbb{E}(\alpha X+\beta)=\alpha\mathbb{E}(X)+\beta$ • $\mathbb{E}(\alpha X+\beta Y)=\alpha\mathbb{E}(X)+\beta\mathbb{E}(Y)$ • For orthogonal variables: ○ $\mathbb{E}(XY)=0$ ○ $\mathbb{E}((X+Y)^2)=\mathbb{E}(X^2)+\mathbb{E}(Y^2)$ • For independent variables: ○ $\mathbb{E}(X|Y)=\mathbb{E}(X)$ ○ $\mathbb{E}(XY)=\mathbb{E}(X)\mathbb{E}(Y)$ • For vectors: If $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$: $\mathbb{E}(\boldsymbol{A}\boldsymbol{x})=\boldsymbol{A}\mathbb{E}(\boldsymbol{x})$ • $\mathbb{E}[\mathbb{E}(X|A)]=\mathbb{E}(X)$ • *Cauchy Schwarz ineq.*: $\mathbb{E}(XY)^2\leq\mathbb{E}(X^2)\mathbb{E}(Y^2)$

Median — Real number M defined by $P(X<M)=P(X>M)$

Standard deviation — $\sqrt{\mathbb{V}(X)}$

Covariance — • Univariate variance of a random variable: $\mathbb{V}(X)=\mathbb{E}((X-\mathbb{E}(X))^2)=\mathbb{E}(X^2)-\mathbb{E}(X)^2$ where $\mathbb{E}(X^2)$ is the unnormalized correlation resp. inner product • Univariate covariance of two random variables: $\text{Cov}(X,Y)=\mathbb{E}((X-\mathbb{E}(X))(Y-\mathbb{E}(Y)))=\mathbb{E}(XY)-\mu_X\mu_Y$ where $\mathbb{E}(XY)$ is the unnormalized correlation resp. inner product • Proof (schematically for variance): $\mathbb{V}(X)=\mathbb{E}((X-\mathbb{E}(X))^2)=\mathbb{E}[X^2-X\mathbb{E}(X)-X\mathbb{E}(X)+\mathbb{E}(X)^2]=\mathbb{E}[X^2]-\mathbb{E}[X]\mathbb{E}(X)-\mathbb{E}[X]\mathbb{E}(X)+\mathbb{E}(X)^2=\mathbb{E}(X^2)-\mathbb{E}(X)^2$ where $\mathbb{E}(X^2)$ is the second moment • Multivariate covariance matrix of a vector: ○ $\boldsymbol{\Sigma}=\text{Cov}(\boldsymbol{x})=\mathbb{E}((\boldsymbol{x}-\mathbb{E}(\boldsymbol{x}))(\boldsymbol{x}-\mathbb{E}(\boldsymbol{x}))^\top)=\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)-\mathbb{E}(\boldsymbol{x})\mathbb{E}(\boldsymbol{x})^\top=\boldsymbol{R}-\boldsymbol{\mu}_X\boldsymbol{\mu}_X^\top=$ $\begin{bmatrix}\mathbb{E}(x_1^2)|\mathbb{E}(x_1x_m)\\ \mathbb{E}(x_mx_1)|\mathbb{E}(x_m^2)\end{bmatrix}-\begin{bmatrix}\mathbb{E}(x_1)^2|\mathbb{E}(x_1)\mathbb{E}(x_m)\\ \mathbb{E}(x_m)\mathbb{E}(x_1)|\mathbb{E}(x_m)^2\end{bmatrix}=\begin{bmatrix}\mathbb{V}(x_1)|\text{Cov}(x_1,x_m)\\ \text{Cov}(x_m,x_1)|\mathbb{V}(x_m)\end{bmatrix}$ where $\boldsymbol{R}=\mathbb{E}(\boldsymbol{x}\boldsymbol{x}^\top)$ is the unnormalized correlation matrix ○ $\boldsymbol{\Sigma}$ and $\boldsymbol{R}$ are symmetric and psd

Properties - variance: • $\mathbb{V}(\alpha)=0$ • $\mathbb{V}(\alpha X+\beta)=\alpha^2\mathbb{V}(X)$ • $\mathbb{V}(X+Y)=\mathbb{V}(X)+2\text{Cov}(X,Y)+\mathbb{V}(Y)$ • For uncorrelated (and independent) variables: $\mathbb{V}(X+Y)=\mathbb{V}(X)+\mathbb{V}(Y)$ • For independent variables: $\mathbb{V}(XY)=\mathbb{E}((XY)^2)\mathbb{E}(XY)^2$ • For vector $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$: $\mathbb{V}_y=\boldsymbol{A}\mathbb{V}_X\boldsymbol{A}^\top$ • For zero-mean variable: $\mathbb{V}(X)=\mathbb{E}(X^2)-\mathbb{E}(X)^2=\mathbb{E}(X^2)$ since $\mathbb{E}(X)=0$ • $\mathbb{V}[X]=\mathbb{V}(\mathbb{E}[X|Y])+\mathbb{E}[\mathbb{V}[X|Y]]$ • $\mathbb{V}[\sum_i x_i]=\sum_i\sum_j\text{Cov}[x_i,x_j]=\sum_i\mathbb{V}[x_i]+\sum_{i\neq j}\text{Cov}[x_i,x_j]$

Properties - covariance: • $\text{Cov}(X,X)=\mathbb{V}(X)$ • $\text{Cov}((\alpha X+\beta Y),Z)=\alpha\text{Cov}(X,Z)+\beta\text{Cov}(Y,Z)$ • If covariance of 2 random variables is 0 resp. $\mathbb{E}(XY)=\mathbb{E}(X)\mathbb{E}(Y)$, they are uncorrelated, but not necessarily independent • If unnormalized correlation of 2 random variables is 0 resp. $\mathbb{E}(XY)=0$, they are orthogonal, but not necessarily independent • For vector $\boldsymbol{y}=\boldsymbol{A}\boldsymbol{x}$: ○ $\boldsymbol{\Sigma}_y=\boldsymbol{A}\boldsymbol{\Sigma}_X\boldsymbol{A}^\top$ ○ $\boldsymbol{R}_y=\boldsymbol{A}\boldsymbol{R}_X\boldsymbol{A}^\top$ • For zero-mean variables: $\text{Cov}(X,Y)=\mathbb{E}(XY)-\mu_X\mu_Y=\mathbb{E}(X,Y)$ since $\mu_X=\mu_Y=0$ • *Cauchy Schwarz ineq.*: $\text{Cov}(X,Y)^2\leq\mathbb{V}(X)\mathbb{V}(Y)$

Correlation — Normalized covariance • Univariate correlation of a random variable: $\text{Cor}(X,Y)=\frac{\text{Cov}(X,Y)}{\sqrt{\mathbb{V}(X)}\sqrt{\mathbb{V}(Y)}}$ • Multivariate correlation matrix of a vector: • $\boldsymbol{P}=\text{Cor}(X)$ ○ $\boldsymbol{P}$ is symmetric and psd ○ Correlation is bounded between 0 and 1, given Cauchy Schwarz Inequality ○ If correlation of two random variables is 0, they are not necessarily independent

Probability Distributions *Normal distribution* — $X\sim N(\mu,\sigma^2)$

For univariate, PDF: $\frac{1}{\sigma\sqrt{2\pi}}exp(\frac{-(x-\mu)^2}{2\sigma^2})=\frac{1}{\sigma\sqrt{2\pi}}exp(-x^2\frac{1}{2\sigma^2}+2x\frac{\mu}{2\sigma^2}-\frac{\mu^2}{2\sigma^2})$ For multivariate, PDF: $\frac{1}{2\pi^{n/2}}\frac{1}{|\Sigma|^{1/2}}exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$

Convolution: $\int N(a;Bc,D)\times N(c;e,F)dc=\int N(a;Be,D+BFB^\top$

Standard normal distribution — Normal distribution, standardized via z-score $z=\frac{x-\mu}{\sigma}$, which results in $\mu=0$ and $\sigma=1$

Bernoulli distribution — trial with success (probability p) or failure (probability $1-p$) • $X\sim$Bernoulli(p) • PDF: $p(x)p^x(1-p)^{1-x}$ • Mean: $\mathbb{E}(x)=p$ • Variance: $\mathbb{V}(x)=p(1-p)$ *Binomial distribution* — n independent Bernoulli trials with k successes • $X\sim$Bin(n,p) • PDF: $\binom{n}{k}p^k(1-p)^{n-k}$ • Mean: $\mathbb{E}(x)=np$ • Variance: $\mathbb{V}(x)=np(1-p)$ *Poisson distribution* — • $X\sim$Pois(λ) • PDF: $e^{-\lambda}\frac{\lambda^x}{x!}$ • Mean: $\mathbb{E}(x)=\lambda$ • Variance: $\mathbb{V}(x)=\lambda$ *Beta distribution* — • X takes values $\in[0,1]$ • Represents the probability of a Bernoulli process after observing $\alpha-1$ successes and $\beta-1$ failures • $X\sim$Beta(α,β) where $\alpha,\beta>0$ • PDF: $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ where $\Gamma(\alpha)=\int_0^\infty u^{\alpha-1}e^{-u}du$ • Mean: $\mathbb{E}(x)=\frac{\alpha}{\alpha+\beta}$ • Variance: $\mathbb{V}(x)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ *Dirichlet distribution* — • X takes values $\in[0,1]$ • Multivariate extension of Beta distribution • $Dir(\boldsymbol{x}|\boldsymbol{\alpha})=\frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^n u_k^{\alpha_k-1}$, where $B(\boldsymbol{\alpha})$ is the multivariate generalization of the Beta funct.: $B(\boldsymbol{\alpha})=\frac{\prod_{k=1}^n\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^n\alpha_k)}$ *Uniform distribution* — • Assume x is uniformly distributed between $[a,b]$ • PDF: $f(x)=\frac{1}{b-a}$ if $a\leq x\leq b$, else 0 • CDF: $F(x)=\frac{x-a}{b-a}$ if $a\leq x\leq b$, 1 if $x>b$, else 0

Other Concepts *Law of large numbers* — Sample mean of iid variables converges to population mean as $n\to\infty$ • *Weak law of large numbers*: $\lim_{n\to\infty}P\left(\left|m_X-\frac{1}{n}\sum_{k=1}^n X_k\right|<\epsilon\right)=1$ • *Strong law of large numbers*: $\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^n X_k=m_X$ with probability 1

Union bound — $P(\bigcup_i A_i)\leq\sum_i P(A_i)$

Jensen's ineq. — Relates expected value of a convex funct. of a random variable to the convex funct. of the expected value of that random variable $\mathbb{E}(f(X))\geq f(\mathbb{E}(X))$

Markov's ineq. — $P(x\geq t)\leq\frac{\mathbb{E}(x)}{t}$. Interesting only for $t\geq\mathbb{E}(x)$ because $p(x\geq t)$ must then be less than or equal to 1. Generalizations: • $p(|x|\geq t)\leq\frac{\mathbb{E}(g(|x|))}{g(t)}$ • $p(|x|\geq t)\leq\frac{\mathbb{E}(|x|^n)}{t^n}$

Hoeffding's Lemma — For random variable with $\mathbb{E}[x]=0$, and $a\leq x\leq b$, and $s>0$: $\mathbb{E}[\exp(sx)]\leq\exp(s^2(b-a)^2/8)$

Hoeffding's Inequality — For random variables x_i that fall in the interval $[a_i,b_i]$ with probability 1, and $S_n=\sum_{i=1}^n x_i$, and $t>0$: • $P(Sn-\mathbb{E}_X Sn\geq t)\leq\exp(-\frac{2t^2}{\sum_{i=1}^n(b_i-a_i)^2})$ • $P(Sn-\mathbb{E}_X Sn\leq -t)\leq\exp(-\frac{2t^2}{\sum_{i=1}^n(b_i-a_i)^2})$ Proof: • Consider the probability $P(Sn-\mathbb{E}[Sn]\geq t)$ • Using Markov's ineq.: $P(Sn-\mathbb{E}[Sn]\geq t)=P(e^{s(Sn-\mathbb{E}[Sn])}\geq e^{st})\leq\frac{\mathbb{E}[e^{s(Sn-\mathbb{E}[Sn])}]}{e^{st}}$ • Using independence of $X_1,...,X_n$: $\mathbb{E}[e^{s(Sn-\mathbb{E}[Sn])}]=\prod_{i=1}^n\mathbb{E}[e^{s(X_i-\mathbb{E}[X_i])}]$ • For each term, we use the fact that $X_i\in[a_i,b_i]$, and apply the lemma ineq.: $\prod_{i=1}^n\mathbb{E}[e^{s(X_i-\mathbb{E}[X_i])}]\leq\prod_{i=1}^n\exp(\frac{s^2(b_i-a_i)^2}{8})$ • Plugging this back in: $e^{-st}\times\mathbb{E}[e^{s(Sn-\mathbb{E}[Sn])}]\leq e^{-st}\times\prod_{i=1}^n\exp(\frac{s^2(b_i-a_i)^2}{8})=e^{-st}\times\exp(\frac{s^2}{8}\sum_{i=1}^n(b_i-a_i)^2)=P(Sn-\mathbb{E}[Sn]\geq t)\leq\exp(-st+\frac{s^2}{8}\sum_{i=1}^n(b_i-a_i)^2)$ • If we set $s=\frac{4t}{\sum_{i=1}^n(b_i-a_i)^2}$ to min. the bound, we get: $P(Sn-\mathbb{E}[Sn]\geq t)\leq\exp(-\frac{2t^2}{\sum_{i=1}^n(b_i-a_i)^2})$ • Similarly for $P(Sn-\mathbb{E}[Sn]\leq -t)$. In the special case of normalized sums of iid variables, where $\tilde{S}=Sn/n$ and $t=n\epsilon$: • Delta given by: ○ $P(\tilde{S}n-\mathbb{E}_X\tilde{S}n\geq\epsilon)\leq\exp(-\frac{2n\epsilon^2}{\sum_{i=1}^n(b_i-a_i)^2/n})$ ○ As $n\to\infty$, this $P(\tilde{S}n-\mathbb{E}_X\tilde{S}n\geq\epsilon)\to 0$ • Absolute deviation given by: ○ $P(|\tilde{S}n-\mathbb{E}_X\tilde{S}n|\geq\epsilon)=P(\tilde{S}n-\mathbb{E}_X\tilde{S}n\geq\epsilon\vee\tilde{S}n-\mathbb{E}_X\tilde{S}n\leq-\epsilon)$ ○ By the union bound: $=P(\tilde{S}n-\mathbb{E}_X\tilde{S}n\geq\epsilon)+P(\tilde{S}n-\mathbb{E}_X\tilde{S}n\leq-\epsilon)$ ○ By Hoeffding's ineq.: $\leq 2\exp(-\frac{2n\epsilon^2}{\sum_{i=1}^n(b_i-a_i)^2/n})$

Chebychev's ineq. — $p(|x-\mu_X|\geq\alpha|\sigma_X|)\leq\frac{1}{\alpha^2}$ resp. $p(|x-\mu_X|\geq\alpha)\leq\frac{|\sigma_X|}{\alpha^2}$. Interesting only for $\alpha>1$. Implications: • For n variables: $p(|Sn-\mu_X|\geq\epsilon)\leq\frac{\sigma_X^2}{n\epsilon^2}$ where $Sn=\frac{1}{n}\sum_{k=1}^n X_k$ is the sample mean

Sufficient statistics — • $Z=g(Y)$ is a sufficient statistic for estimating X if X can be estimated as well from Z as from Y, i.e. condensing Y to Z does not entail any loss of information about X • Conditioned on Z, Y is independent of X: $p(Y|Z,X)=p(Y|Z)$ • For sufficient statistics, the MLE of X from Y is the same as the MLE of X from Z: $argmax_X p(Y|X)\rho(y)=argmax_X p(Z|X)\rho(y)$ • $p(X|Z)=p(X|Y)$

Hypothesis Testing *Terminology* — • Hypothesis: ○ H_0: Accepted null hypothesis, e.g. $p=p_0$, $p_1-p_2=p_{0,1}-p_{0,2}=0$ ○ H_A: Alternative hypothesis, e.g. $p\neq p_0$, $p_1-p_2\neq p_{0,1}-p_{0,2}\neq 0$ • Test types: ○ *Two-sided*: $H_0:p=p_0,H_A:p\neq p_0$ ○ *One-sided upper tail*: $H_0:p\leq p_0,H_A:p>p_0$ ○ *One-sided lower tail*: $H_0:p\geq p_0,H_A:p<p_0$ • Errors: ○ *True positive*: Chose H_0, and H_0 obtains ○ *False positive, type I error*: Chose H_A, but H_0 obtains ○ *True negative*: Chose H_A, and H_A obtains ○ *False negative, type II error*: Chose H_0, but H_A obtains • *Significance level* α: ○ $\alpha\geq p$(type I error)$=p(\bar{x}\geq c|H_0)=p(z_n\geq z_\alpha|H_0)$ with eq. for continuous variables ○ If α is small, the probability that we are erroneously rejecting H_0 is very small ○ Set by us, typically at 5% • *Critical value* z: ○ For two-sided: $z_{\alpha/2}$, $z_{1-\alpha/2}$ ○ For one-sided upper tail: $z_{1-\alpha}$ ○ For one-sided lower tail: z_α ○ Associated z-score with α ○ Corresponds to threshold c prior to z-score transformation • *P-value* p: ○ For two-sided: $p=P(|z|\geq z_n)$ ○ For one-sided upper tail: $p=P(z\geq z_n)$ ○ For one-sided lower tail: $p=P(z\leq z_n)$ ○ Probability, given H_0 that we observe a value as or more extreme as the observed value z_n ○ Smallest significance level resp. largest confidence level, at which we can reject H_0 given the sample observed ○ If p-value is less than significance level α resp. if observed value z_n is more extreme than critical value z_α resp. if observed mean $\bar{x}$ is more extreme than threshold c, reject H_0, because the probability that we are erroneously doing so is very small • *Confidence level*: $1-\alpha$, probability, given H_0, that we retain H_0 • *Beta*: $\beta=p$(type II error) • *Power*: ○ $1-\beta=p(\bar{x}\geq c|H_1)=p(z_n\geq z_\alpha|H_1)$ ○ Probability, given H_A, that we reject H_0 • *Test statistic*: $z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$, $z|H_0\sim N(0,1)$, $z|H_1\sim N(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}},1)$ • Rejection region: $\bar{x}>\mu_0+\sigma z_\alpha/\sqrt{n}$ • Example: ○ If $X\sim N(\theta,1)$ and $H_0:\theta=0$: ○ $\alpha=p(\bar{x}\geq c|H_0)=p(\sqrt{n}\bar{x}\geq\sqrt{n}c|H_0)=p(z_n\geq\sqrt{n}c|H_0)=1-\Phi(\sqrt{n}c)$ where ■ Φ is the CDF of the normal distribution ■ $z_n|H_0=\frac{\bar{x}-0}{1/\sqrt{n}}=\sqrt{n}\bar{x}$ ○ $c=\frac{1}{\sqrt{n}}\Phi^{-1}(1-\alpha)$ ○ $1-\beta=p(\bar{x}\geq c|H_1)=p(\sqrt{n}(\bar{x}-1)\geq\sqrt{n}(c-1)|H_1)=p(z_n\geq\sqrt{n}(c-1)|H_1)=p(z_n\geq\sqrt{n}(c-1)|H_0)=1-\Phi(\sqrt{n}(c-1))$ where ■ Φ is the CDF of the normal distribution ■ $z_n|H_1=\frac{\bar{x}-1}{1/\sqrt{n}}=\sqrt{n}(\bar{x}-1)$ ■ We can switch from $|H_1$ to $|H_2$ because the two distributions follow the same form, just shifted and $1-\beta=p(z_n\geq z_\alpha|H_1)=p(z_n\geq z_\alpha-\delta|H_0)=1-\Phi(z_\alpha-\delta)$ where $\delta=\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}$

Multiple comparisons problem — Accumulation of false positive rate (α) for K tests, due to independence of tests: P(|false rejections of H_0|>0)$=1-P$(|false rejections of H_0|=0)$=1-(1-\alpha)^K$

Corrections for multiple comparisons problem — *Bonferroni correction*: New significance level set to $\alpha*=\alpha/K$

Neyman Pearson test — • Max. power while controlling type I errors • Sets α such that $\alpha\geq p$(type I error) • Then minimizes p(type II error) • This is achieved by a likelihood-ratio test with threshold θ, such that α equals or is as close as possible to p(type I error): ○ If $\Lambda(x)=\frac{p(x|p_0)}{p(x|p_A)}>\theta$, we reject H_0 ○ Then, we have $P(\Lambda(x)>\theta|H_0)=P(\frac{p(x|p_0)}{p(x|p_A)}>\theta|H_0)=P$(type I error)$=\alpha$ ○ The smaller α, the larger θ

ML Decision Rule — • If $L(x)=\log(\frac{p(x|p_A)}{p(x|p_0)})>0$, we reject H_0 *MAP Decision Rule* — • If $M(x)=\log(\frac{p(x|p_A)P(p_A)}{p(x|p_0)P(p_0)})>0$, we reject H_0 *Bayesian Hypothesis Testing* — • If $\Lambda(x)=\frac{p(x|p_A)}{p(x|p_0)}>\theta$, we reject H_0 • $\theta=\frac{k(p_A,p_0)P(p_0)}{k(p_0,p_A)P(p_A)}$ • In this case, θ subsumes both the prior $p(x)$ and the costs $k(\hat{x},x)$ *Error probability* — • $p(p_{\text{not chosen}}|x)$

Information Theory *Entropy* — • $H(x)=\mathbb{E}[-\log p(x)]=-\sum_x p(x)\log(p(x))$ Properties: • $H\geq 0$ • H is maximized, when x is a uniform random variable • For independent variables: $H(x,y)=H(x)+H(y)$ *Conditional entropy* — • $H(x|y)=-\sum_{x,y}p(y)p(x|y)log(p(x|y))=-\sum_{x,y}p(x,y)log(\frac{p(x,y)}{p(y)})$ Properties: • $0\leq H(x|y)\leq H(x)$ with eq. if when x is independent with y resp. if y completely determines x *Mutual information* — • $I(x;y)=\mathbb{E}[\log(p(x,y)/p(y)p(x))]=\sum_{x,y}p(x,y)\log(p(x,y)/p(x)p(y))$ • $I(x;y)=H(x)-H(x|y)$ Properties: ○ $0\leq I(x;y)\leq H(x)$ with eq. if y completely determines x resp. if x is independent with y *KL divergence* — • $KL(x;y)=\mathbb{E}[\log(p(x)/q(x))]=\sum_x p(x)\log(p(x)/q(x))=\sum_x p(x)\log(p(x))-\sum_x p(x)\log(q(x))=\sum_x p(x)\log(p(x))-1\times$const. if q(x) unif. • $KL(x;y)=-H(p)-\sum_x p(x)\log(q(x))$ Properties: • $KL(x;y)\geq 0$ *Cross entropy* — • $CE(p|q)=KL(p;q)+H(p)=-\sum_x p(x)log(q(x))$ *ELBO* — ELBO$(z)=\mathbb{E}[\log(p(x,z)/q(z))]=\sum_z q(z)\log(p(x,z)/q(z))$

ML Paradigms

Frequentism *Descr.* — • Parametric approach • $\boldsymbol{\theta}$ as fixed, unknown quantity, $\boldsymbol{X}$ as random, and known quantity • Makes point est. • Focuses on max. likelihood $p(\boldsymbol{X}|\boldsymbol{\theta})$ to infer posterior $p(\boldsymbol{\theta}|\boldsymbol{X})$ • Only requires differentiation methods • High variance, but low bias

MLE est. • Max. log-likelihood: $\hat{\boldsymbol{\theta}}=\arg\max_{\boldsymbol{\theta}}(L)=\arg\max_{\boldsymbol{\theta}}(p(y_1,...,y_n|\boldsymbol{x}_i,\boldsymbol{\theta}))=\arg\max_{\boldsymbol{\theta}}(\prod_{i=1}^n p(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}))=\arg\max_{\boldsymbol{\theta}}(\sum_{i=1}^n log(p(y_i|\boldsymbol{x}_i,\boldsymbol{\theta})))$ • In discrete case: ○ $\hat{\boldsymbol{\theta}}=\arg\max_{\boldsymbol{\theta}}(L)=\arg\max_{\boldsymbol{\theta}}(\prod_{i=1}^n p(y_i|\boldsymbol{x}_i,\boldsymbol{\theta}))=\arg\max_{\boldsymbol{\theta}}\prod_{j=1}^k p_j^{N_j}=\arg\max_{\boldsymbol{\theta}}\sum_{i=1}^n N_j log(p_j)$ where ■ $j=1,...,k$ is the number of classes ■ N_j county how often the outcome class j appears in $\boldsymbol{y}$ ■ $p_j=p(y_i=j|\boldsymbol{x}_i,\boldsymbol{\theta})$ ○ We can further expand to $\hat{\boldsymbol{\theta}}=\arg\max_{\boldsymbol{\theta}}(L)=\arg\max_{\boldsymbol{\theta}}\sum_{i=1}^n N_j log(p_j)=\arg\max_{\boldsymbol{\theta}}\sum_{i=1}^n\frac{N_j}{n}log(p_j)=\arg\max_{\boldsymbol{\theta}}\sum_{i=1}^n\frac{N_j}{n}(log(\frac{p_j}{N_j/n})+log(N_j/n)))=\arg\max_{\boldsymbol{\theta}}\sum_{i=1}^n\frac{N_j}{n}log(\frac{p_j}{N_j/n})=\arg\min_{\boldsymbol{\theta}}\sum_{i=1}^n\frac{N_j}{n}log(\frac{N_j/n}{p_j})=\arg\min_{\boldsymbol{\theta}}\sum_{i=1}^n\tilde{p}_j log(\frac{\tilde{p}_j}{p_j})$ ○ This can be solved using constrained opt. with strong duality subject to $\sum_j p_j=1$ ○ We then get $\theta_{MLE}=N_j/n$ which minimizes the KL divergence when $\tilde{p}_j=p_j$ • *Score*: ○ The score is the derivative of the log-likelihood: $\Lambda=\frac{\partial}{\partial\boldsymbol{\theta}}log(p(y|\boldsymbol{x},\boldsymbol{\theta}))=\frac{\frac{\partial}{\partial\boldsymbol{\theta}}p(y|\boldsymbol{x},\boldsymbol{\theta})}{p(y|\boldsymbol{x},\boldsymbol{\theta})}$ ○ The expected score is given by: $\mathbb{E}(\Lambda)=\int p(y|\boldsymbol{x},\boldsymbol{\theta})\frac{\frac{\partial}{\partial\boldsymbol{\theta}}p(y|\boldsymbol{x},\boldsymbol{\theta})}{p(y|\boldsymbol{x},\boldsymbol{\theta})}dx=\frac{\partial}{\partial\boldsymbol{\theta}}\int p(y|\boldsymbol{x},\boldsymbol{\theta})dx=\frac{\partial}{\partial\boldsymbol{\theta}}\times 1=0$ • Equivalent to min. KL divergence between as $n\to\infty$: ○ $\hat{\theta}_{MLE}=\arg\min_\theta\prod_{i=1}^n q(x_i|\theta)=\sum_{i=1}^n\log q(x_i|\theta)$ ○ $\hat{\theta}_{KL}=\arg\min_\theta D_{KL}(\hat{p}(x)\|q(x|\theta))$ ○ $D_{KL}(\hat{p}(x)\|q(x|\theta))=\sum_{i=1}^n\hat{p}(x_i)\log\frac{\hat{p}(x_i)}{q(x_i|\theta)}=\frac{1}{n}\sum_{i=1}^n\log\frac{\hat{p}(x_i)}{q(x_i|\theta)}=\frac{1}{n}\sum_{i=1}^n\log\hat{p}(x_i)-\frac{1}{n}\sum_{i=1}^n\log q(x_i|\theta)$ ○ Min. D_{KL} is equivalent to max.: $\frac{1}{n}\sum_{i=1}^n\log q(x_i|\theta)$

Bayesianism *Descr.* — • Parametric approach • $\boldsymbol{\theta}$ as random, unknown quantity, $\boldsymbol{X}$ as random and known quantity • Makes est. in form of distribution • Leverages prior and likelihood to infer posterior: $p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y})=\frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})}{p(\boldsymbol{y}|\boldsymbol{X})}=\frac{p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})d\boldsymbol{\theta}}\propto p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})=p(\boldsymbol{\theta},\boldsymbol{y}|\boldsymbol{X})$ • Focuses on min. cost funct. $\mathbb{E}[k(\boldsymbol{\theta}',\boldsymbol{\Theta})|\boldsymbol{X},\boldsymbol{y}]=\int_{\boldsymbol{\theta}}p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y})\times k(\boldsymbol{\theta}',\boldsymbol{\theta})d\boldsymbol{\theta}\propto\int_{\boldsymbol{\theta}}p(\boldsymbol{\theta},\boldsymbol{y}|\boldsymbol{X})\times k(\boldsymbol{\theta}',\boldsymbol{\theta})d\boldsymbol{\theta}$ resp. $\sum p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y})\times k(\boldsymbol{\theta}',\boldsymbol{\theta})$ • Requires integration methods for normalizing constant in denominator, which can be intractable • Low variance, but high bias

MMSE est. • Min. MSE as cost funct. $k(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})=|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}|^2$ • The resulting est. is the mean of the posterior: $\hat{\boldsymbol{\theta}}=\mathbb{E}[\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y}]$. Proof: ○ $\mathbb{E}[|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}|^2|\boldsymbol{y}]=\int(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^2p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^2-2\hat{\boldsymbol{\theta}}\int\boldsymbol{\theta}p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}+\int\boldsymbol{\theta}^2p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$ ○ Taking the derivative with respect to $\hat{\boldsymbol{\theta}}$: $\frac{\partial\mathbb{E}[|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}|^2\boldsymbol{y}]}{\partial\hat{\boldsymbol{\theta}}}=2\hat{\boldsymbol{\theta}}-2\int\boldsymbol{\theta}p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$ ○ Setting the derivative to zero: $\hat{\boldsymbol{\theta}}=\int\boldsymbol{\theta}p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}=\mathbb{E}[\boldsymbol{\theta}|\boldsymbol{y}]$ • Returns single point est.

Median est. • Min. mean absolute error as cost funct. $k(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})=|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}|$ • The resulting est. is the median of the posterior. Proof: ○ Bayesian cost funct. given by: $\mathbb{E}[|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}||\boldsymbol{y}]=\int|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}|p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$ ○ The integral splits into two parts: $=\int_{-\infty}^{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}+\int_{\hat{\boldsymbol{\theta}}}^{\infty}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$ ○ Taking the derivative with respect to $\hat{\boldsymbol{\theta}}$: $\frac{\partial\mathbb{E}[|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}||\boldsymbol{y}]}{\partial\hat{\boldsymbol{\theta}}}$ for each term separately ○ $\frac{\partial}{\partial\hat{\boldsymbol{\theta}}}\int_{-\infty}^{\hat{\boldsymbol{\theta}}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}=\int_{-\infty}^{\hat{\boldsymbol{\theta}}}p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}p(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ ○ $\frac{\partial}{\partial\hat{\boldsymbol{\theta}}}\int_{\hat{\boldsymbol{\theta}}}^{\infty}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}=-\int_{\hat{\boldsymbol{\theta}}}^{\infty}p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}p(\hat{\boldsymbol{\theta}}|\boldsymbol{y})$ ○ Combining the two derivatives, we get: $\int_{-\infty}^{\hat{\boldsymbol{\theta}}}p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}-\int_{\hat{\boldsymbol{\theta}}}^{\infty}p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$ ○ Setting the derivative to zero: $\int_{-\infty}^{\hat{\boldsymbol{\theta}}}p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}=\int_{\hat{\boldsymbol{\theta}}}^{\infty}p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$ ○ Since the total probability is 1, this implies: $\int_{-\infty}^{\hat{\boldsymbol{\theta}}}p(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}=0.5$ • Returns single point est.

MAP est. • Max. posterior: $\hat{\boldsymbol{\theta}}=\arg\max_{\boldsymbol{\theta}}p(\boldsymbol{\theta}|\boldsymbol{X})\propto\arg\max_{\boldsymbol{\theta}}p(\boldsymbol{\theta}|\boldsymbol{X})p(\boldsymbol{X})$ • The resulting est. is the mode of the posterior • Returns single point est. • In discrete case: ○ MAP minimizes zero-one loss as cost funct.: $k(\hat{\boldsymbol{\theta}},\boldsymbol{\theta})=\begin{cases}1 & \hat{\boldsymbol{\theta}}\neq\boldsymbol{\theta}\\0 & \hat{\boldsymbol{\theta}}=\boldsymbol{\theta}\end{cases}$ ○ We can make $\theta_{MLE}=N_j/n$ more robust by setting a prior $p(\boldsymbol{\theta})\propto\prod_{j=1}^n p_j^\nu$ with parameter $0<\nu\leq 1$ ○ $\hat{\boldsymbol{\theta}}=\arg\max_{\boldsymbol{\theta}}(p(\boldsymbol{\theta}|\boldsymbol{y}))=\arg\max_{\boldsymbol{\theta}}(p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}))=\arg\max_{\boldsymbol{\theta}}\prod_{j=1}^k p_j^{N_j}\prod_{j=1}^k p_j^\nu=\arg\max_{\boldsymbol{\theta}}\prod_{j=1}^k p_j^{N_j+\nu}=\arg\max_{\boldsymbol{\theta}}\sum_{j=1}^k(N_j+\nu)\log(p_j)=\arg\max_{\boldsymbol{\theta}}\sum_{j=1}^k\frac{N_j+\nu}{n+k\nu}\log(\frac{p_j}{(N_j+\nu)/(n+k\nu)})=\arg\min_{\boldsymbol{\theta}}\sum_{j=1}^k\frac{N_j+\nu}{n+k\nu}\log(\frac{(N_j+\nu)/(n+k\nu)}{p_j})=\arg\min_{\boldsymbol{\theta}}\sum_{j=1}^k\tilde{p}_j\log(\frac{\tilde{p}_j}{p_j})$ ○ This can be solved using constrained opt. with strong duality subject to $\sum_j p_j=1$ ○ We then get $\theta_{MAP}=(N_j+\nu)/(n+k\nu)$ which minimizes the KL divergence when $\tilde{p}_j=p_j$

Model Opt.

Gradient Descent $\boldsymbol{\beta}_{(t+1)}\leftarrow\boldsymbol{\beta}_{(t)}-\eta\nabla_{\boldsymbol{\beta}}LO|_{\boldsymbol{\beta}=\boldsymbol{\beta}_{(t)}}$

Model Evaluation

Estimator Evaluation Criteria *Criteria* — • Consistency: $\hat{\boldsymbol{\theta}}\to\boldsymbol{\theta}$ as $n\to\infty$ • Bias: $\mathbb{E}(\hat{\boldsymbol{\theta}})-\boldsymbol{\theta}$ • Unbiased: $\mathbb{E}(\hat{\boldsymbol{\theta}})=\boldsymbol{\theta}$ • Asymptotically unbiased: $\mathbb{E}[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^2]=0$ as $n\to\infty$

Bias Variance Tradeoff • Mean squared error $\mathbb{E}[(\hat{f}(\boldsymbol{X})-\boldsymbol{y})^2]$ can be decomposed into: $(\mathbb{E}[\hat{f}(\boldsymbol{X})]-f(\boldsymbol{X}))^2+\mathbb{V}(\hat{f}(\boldsymbol{X}))+\mathbb{E}[\boldsymbol{\epsilon}^2]=$ bias^2+variance+irreducible error. Proof: ○ $\boldsymbol{y}=f(\boldsymbol{X})+\boldsymbol{\epsilon}$ ○ $\mathbb{E}[(\hat{f}(\boldsymbol{X})-\boldsymbol{y})^2]=\mathbb{E}[(\hat{f}(\boldsymbol{X})-\mathbb{E}[\hat{f}(\boldsymbol{X})]+\mathbb{E}[\hat{f}(\boldsymbol{X})]-f(\boldsymbol{X})+\boldsymbol{\epsilon})^2]=\mathbb{E}[(\hat{f}(\boldsymbol{X})-\mathbb{E}[\hat{f}(\boldsymbol{X})])^2]+\mathbb{E}[(\mathbb{E}[\hat{f}(\boldsymbol{X})]-f(\boldsymbol{X}))^2]+\mathbb{E}[\boldsymbol{\epsilon}^2]-2\mathbb{E}[(\hat{f}(\boldsymbol{X})-\mathbb{E}[\hat{f}(\boldsymbol{X})])(\mathbb{E}[\hat{f}(\boldsymbol{X})]-f(\boldsymbol{X})+\boldsymbol{\epsilon})]$ ○ Third term $\mathbb{E}[\boldsymbol{\epsilon}^2]$ is the variance of $\boldsymbol{y}$: $=\mathbb{E}[\boldsymbol{\epsilon}^2]-\mathbb{E}[\boldsymbol{\epsilon}]^2=\mathbb{V}(\boldsymbol{y})=\sigma^2$ ○ Fourth term $2\mathbb{E}[(\hat{f}(\boldsymbol{X})-\mathbb{E}[\hat{f}(\boldsymbol{X})])(\mathbb{E}[\hat{f}(\boldsymbol{X})]-f(\boldsymbol{X})+\boldsymbol{\epsilon})]$ equals 0: ■ $2\mathbb{E}[(\hat{f}(\boldsymbol{X})-\mathbb{E}[\hat{f}(\boldsymbol{X})])(\mathbb{E}[\hat{f}(\boldsymbol{X})]-f(\boldsymbol{X})+\boldsymbol{\epsilon})]=2(\mathbb{E}[\hat{f}(\boldsymbol{X})]-f(\boldsymbol{X})+\boldsymbol{\epsilon}\mathbb{E}[(\hat{f}(\boldsymbol{X})-\mathbb{E}[\hat{f}(\boldsymbol{X})])]$ because $(\mathbb{E}[\hat{f}(\boldsymbol{X})]-f(\boldsymbol{X})+\boldsymbol{\epsilon})$ is deterministic ■ In last equation, second term equals 0, so whole equation is 0 ○ Then, we are left with: variance+bias^2+irreducible error

Estimating Common Distributions

Gaussian *Frequentism — MLE* — • Likelihood (excl. constants): $L=(\frac{1}{\sigma})^n\prod_{i=1}^n exp(-\frac{1}{2\sigma^2}(x^{(i)}-\mu)^2)=\frac{1}{\sigma^n}exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x^{(i)}-\mu)^2)=\frac{1}{\sigma^n}exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x^{(i)}-\bar{x}+\bar{x}-\mu)^2)=\frac{1}{\sigma^n}exp(-\frac{\sum_{i=1}^n(x^{(i)}-\bar{x})^2}{2\sigma^2})exp(-\frac{n(\bar{x}-\mu)^2}{2\sigma^2})=\frac{1}{\sigma^n}exp(-\frac{nS^2}{2\sigma^2})exp(-\frac{n(\bar{x}-\mu)^2}{2\sigma^2})$ where $S=\frac{1}{n}\sum(x^{(i)}-\bar{x})^2$ is the covariance matrix • Log-likelihood: $LL=-nlog(\sigma)-\sum_{i=1}^n(\frac{1}{2\sigma^2}(x^{(i)}-\mu)^2)=-nlog(\sigma)-\frac{nS^2}{2\sigma^2}-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}$ • $\boldsymbol{\mu}_{MLE}$ is sample mean: $\frac{1}{n}\sum_{i=1}^n x^{(i)}$: ○ Derivative of log-likelihood wrt $\boldsymbol{\mu}$: $\nabla_{\boldsymbol{\mu}}LL=\nabla_{\boldsymbol{\mu}}(-\sum_{i=1}^n(\frac{x^{(i)2}-2x^{(i)}\boldsymbol{\mu}+\boldsymbol{\mu}^2}{2\sigma^2}))=\nabla_{\boldsymbol{\mu}}(-\sum_{i=1}^n(-\frac{x^{(i)}\boldsymbol{\mu}}{\sigma^2}+\frac{\boldsymbol{\mu}^2}{2\sigma^2}))=-\sum_{i=1}^n(-\frac{x^{(i)}}{\sigma^2}+\frac{2\boldsymbol{\mu}}{2\sigma^2})=\sum_{i=1}^n(\frac{x^{(i)}-\boldsymbol{\mu}}{\sigma^2})=\sum_{i=1}^n x^{(i)}-n\mu=0$ • $\boldsymbol{\mu}_{MLE}$ is an unbiased est.: ○ $\mathbb{E}[\boldsymbol{\mu}_{MLE}]=\mathbb{E}[\frac{1}{n}\sum_{i=1}^n x_i]=\frac{1}{n}\sum_{i=1}^n\mathbb{E}[x_i]=\boldsymbol{\mu}$ • $\boldsymbol{\sigma}^2_{MLE}$ is sample variance: $\frac{1}{n}\sum_{i=1}^n(x^{(i)}-\hat{\mu})^2=\frac{1}{n}\sum_{i=1}^n x^{(i)}x^{(i)\top}-\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top$: ○ Derivative of log-likelihood wrt $\boldsymbol{\sigma}$: $\nabla_\sigma LL=-n\nabla_\sigma log(\sigma)-\nabla_\sigma(\sum_{i=1}^n(\frac{(x^{(i)}-\hat{\mu})^2}{2\sigma^2}))=\frac{-n}{\sigma}-\nabla_\sigma(\sum_{i=1}^n\frac{1}{2}\sigma^{-2}(x^{(i)}-\hat{\mu})^2)=\frac{-n}{\sigma}-(\sum_{i=1}^n-1\sigma^{-3}(x^{(i)}-\hat{\mu})^2)=-n+\sum_{i=1}^n(\frac{(x^{(i)}-\hat{\mu})^2}{\sigma^2})=0$ • $\boldsymbol{\sigma}^2_{MLE}$ is a biased est.: ○ $\mathbb{E}[\boldsymbol{\Sigma}_{MLE}]=\frac{1}{n}\sum_{i=1}^n\mathbb{E}[(x_i-\hat{\mu})(x_i-\hat{\mu})^\top]=\frac{1}{n}\sum_{i=1}^n\mathbb{E}[x_ix_i^\top]-\frac{1}{n}\sum_{i=1}^n\mathbb{E}[x_i\hat{\mu}^\top]-\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\hat{\mu}x_i^\top]+\mathbb{E}[\hat{\mu}\hat{\mu}^\top]=\frac{1}{n}\sum_{i=1}^n\mathbb{E}[x_ix_i^\top]-\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\mathbb{E}[x_ix_j^\top]-\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\mathbb{E}[x_ix_j^\top]+\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\mathbb{E}[x_ix_j^\top]=\frac{1}{n}\sum_{i=1}^n\mathbb{E}[x_ix_i^\top]-\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\mathbb{E}[x_ix_j^\top]=\frac{1}{n}n(\Sigma+\mu\mu^\top)-\frac{1}{n^2}(n^2\mu\mu^\top+n\Sigma)$. Proof: ■ $\mathbb{E}[x_ix_j^\top]=\delta_{ij}\Sigma+\mu\mu^\top$ where $\delta=1$ if $i=j$. Proof: □ $\Sigma=\mathbb{E}[(x_i-\mu)(x_i-\mu)^\top]=\mathbb{E}[x_ix_i^\top-2x_i\mu^\top+\mu\mu^\top]=\mathbb{E}[x_ix_i^\top]-\mu\mu^\top$ □ For $i\neq j$, covariance is 0: $\mathbb{E}[x_ix_j^\top]=\mathbb{E}[x_i]\mathbb{E}[x_j]=\mu\mu^\top$ ■ ...$+n\Sigma$ since in n cases $\delta=1$ ○ $=\Sigma-\frac{1}{n}\Sigma$ ○ $\mathbb{E}[\Sigma_{MLE}]=\Sigma-\frac{1}{n}\Sigma$ • σ^2_{MLE} est. satisfies: $\mathbb{E}[\sigma^2_{MLE}]=\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]-\mathbb{E}[\hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top]=\frac{n-1}{n}\Sigma$

Bayesianism — • Assume σ^2 is known and $\mu\sim N(\mu_0,\sigma_0^2)$ is the outcome of a random variable • $p(\mu|x,\mu_0,\sigma_0^2)\propto p(x|\mu,\sigma^2)p(\mu|\mu_0,\sigma_0^2)$ • The likelihood is $p(x|\mu,\sigma^2)=\prod_{i=1}^n\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x_i-\mu)^2}{2\sigma^2})$ • The prior is $p(\mu|\mu_0,\sigma_0^2)=\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2})$ • The, the posterior is given by: $p(\mu|x,\mu_0,\sigma_0^2)\propto\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2)$ • Expanding the likelihood term: $\sum_{i=1}^n(x_i-\mu)^2=\sum_{i=1}^n(x_i^2-2\mu x_i+\mu^2)=\sum_{i=1}^n x_i^2-2\mu\sum_{i=1}^n x_i+n\mu^2$ • Expanding the prior term: $(\mu-\mu_0)^2=(\mu^2-2\mu\mu_0+\mu_0^2)$ • This yields: $p(\mu|x,\mu_0,\sigma_0^2)\propto\exp(-\frac{1}{2}(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2})\mu^2+(\frac{\sum_{i=1}^n x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})\mu+$(constant terms)) where the constant terms include terms that do not depend on μ, such as $\sum_{i=1}^n x_i^2$ and μ_0^2 • Based on the parametric form of the Gaussian distribution, this yields $(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2})\mu^2=\frac{1}{2\sigma_n^2}\mu^2$ and $(\frac{\sum_{i=1}^n x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})\mu=\frac{\mu_n}{\sigma_n^2}\mu$ • Solving for σ_n and μ_n: ○ $\mu_n=\frac{\frac{\sum_{i=1}^n x_i}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}=\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}=\frac{\frac{n\bar{x}\sigma_0^2+\mu_0\sigma^2}{\sigma^2\sigma_0^2}}{\frac{n\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}}=\frac{n\bar{x}\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2}$ ○ $\sigma_n=\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}=\frac{1}{\frac{n\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2}}=\frac{\sigma^2\sigma_0^2}{n\sigma_0^2+\sigma^2}$ • Thus, the posterior is $p(\mu|x,\mu_0,\sigma_0^2)\sim N(\mu_n,\sigma_n^2)$ • Conjugate prior: Gaussian

Binomial *Frequentism — MLE —* • Likelihood: $P(\delta|p)\sim p^s(1-p)^{n-s}$ • $\hat{p}_{MLE}=\arg\max(p^s(1-p)^{n-s})=s/n$ *Bayesianism —* • Posterior: $P(p|\delta)\propto p^s(1-p)^{n-s}p^{\alpha-1}(1-p)^{\beta-1}=p^{s+\alpha-1}(1-p)^{n-s+\beta-1}\sim\text{Beta}(s+\alpha,n-s+\beta)$ • Conjugate prior: Beta *Bayesianism — MAP —* • $\hat{p}_{MAP}=\frac{s+\alpha-1}{s+\alpha+n-s+\beta-2}$ • Posterior mean

Poisson *Frequentism — MLE —* • Likelihood: $P(x|\lambda)=\prod_{i=1}^n\frac{\lambda^{x_i}e^{-\lambda}}{x_i!}=\lambda^{\sum_{i=1}^n x_i}\frac{e^{-\lambda n}}{\prod_{i=1}^n x_i!}$ • $\hat{\lambda}_{MLE}=\frac{1}{n}\sum_{i=1}^n x_i$ *Bayesianism —* • Posterior: $P(\lambda|x)\propto\lambda^{\sum_{i=1}^n x_i}e^{-\lambda n}\lambda^{\alpha-1}e^{-\beta\lambda}=\lambda^{\sum_{i=1}^n x_i+\alpha-1}\times e^{-\lambda n}\times e^{-\beta\lambda}=\lambda^{\sum_{i=1}^n x_i+\alpha-1}\times e^{-(n+\beta)\lambda}\sim\text{Gamma}(\alpha+\sum_{i=1}^n x_i,\beta+n)$ • Conjugate prior: Gamma *Bayesianism — MAP —* • $\hat{\lambda}_{MAP}=\frac{\alpha+\sum_{i=1}^n x_i-1}{\beta+n}$ • Posterior mean

Linear Regr.

Form. • $y^{(i)}=\beta\cdot x^{(i)}$ resp. $y=X\beta$ where X contains n rows, each of which represents an instance, and m columns, each of which represents a feature • To incorporate offset, first column of X (i.e. first feature) is set to 1 and first element of β is set to β_0 • $\hat{y}=X\beta$ is a projection of y to the columnspace of X • β lies in the rowspace of X resp. columnspace of $X^\top$

Opt. *Objective funct. — Ordinary least squares est. (OLSE):* Min. MSE: $LO=\frac{1}{n}\sum_{i=1}^n(y^{(i)}-\beta\cdot x^{(i)})^2$ resp. $LO=(y-X\beta)^\top(y-X\beta)$ *Opt. —* • $\nabla_\beta LO=\frac{1}{2}\nabla_\beta((y-\beta\cdot x)^2=(y-\beta\cdot x)x=0$ resp. $\nabla_\beta LO=\nabla_\beta((y-X\beta)^\top(y-X\beta))=\nabla_\beta(\beta^\top X^\top X\beta-2y^\top X\beta)=2X^\top X\beta-2X^\top y=X^\top(X\beta-y)=0$ • $\Rightarrow\beta=(X^\top X)^{-1}X^\top y$ *Distr. —* • $\mathbb{E}[y]=X\mathbb{E}[\beta]+b_0+\mathbb{E}[\epsilon]$ • $\mathbb{V}[y]=X\mathbb{V}[\beta]X^\top+\mathbb{V}[\epsilon]$ • $R[y]=\mathbb{E}[yy^\top]=XR[\beta]X^\top+X\mathbb{E}[\beta]b_0^\top+(X\mathbb{E}[\beta]b_0^\top)^\top+b_0b_0^\top$

Alternatives to OLSE — MLE: • Yields same result as OLSE • The likelihood is: $p(y|\beta,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp(-\frac{1}{2\sigma^2}\|y-X\beta\|^2)$ • The log likelihood is: $\mathcal{L}=-\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\|y-X\beta\|^2$ • We min.: $\|y-X\beta\|^2$ • This is equivalent to OLSE Orthogonality principle: • $\hat{y}=X\beta$ is a projection of y to the columnspace of X • Projection matrix for this is $X(X^\top X)^{-1}X^\top$ • Coefficients β are in the columnspace of $X^\top$ if $(X^\top X)^{-1}$ is invertible • Projection matrix for this is $X^\top(XX^\top)^{-1}X$ • Error ϵ is orthogonal to the columnspace of X • We wish to min. $\|\hat{y}-y\|=\|X\beta-y\|$ by selecting β appropriately • By the orthogonality principle, $x^{[j]}\cdot(\hat{y}-y)=x^{[j]}\cdot(X\beta-y)=0$ where $x^{[j]}$ is the j^{th} column of X resp. $X^\top(\hat{y}-y)=X^\top(X\beta-y)=0\Rightarrow\beta=(X^\top X)^{-1}X^\top y$ • Alternatively, for min. norm sol., β lies in the columnspace of $X^\top$ • Then, we can express β as $X^\top[\alpha_1,\dots,\alpha_n]^\top$ • This yields an equation system $y=XX^\top[\alpha_1,\dots,\alpha_n]^\top$ which can be solved for α_i • On that basis, β can be calculated Pseudo Inverse: • Yields same result as OLSE • Minimum-norm sol. • β minimizes MSE if $\hat{y}=X\beta$ is a projection of y to the columnspace of X • Given matrix projection via SVD, $XX^{\#}y$ is that projection • $\Rightarrow\beta=X^{\#}y$ • If X has full column rank: $=(X^\top X)^{-1}X^\top y$ • If X has full row rank: $=X^\top(XX^\top)^{-1}y$ PCA: • Instances $y^{(i)},x^{(i)}=\xi^{(i)}$ can be projected onto hyperplane given by $X\beta$ • Projections are given by $\hat{\xi}^{(i)}$ • Residuals are given by $e^{(i)}=\xi^{(i)}-\hat{\xi}^{(i)}$ • Since $e^{(i)}$ is orthogonal to $\hat{\xi}^{(i)}$, we can write using Pythagorean theorem: $\|e^{(i)}\|^2=\|\xi^{(i)}\|^2-\|\hat{\xi}^{(i)}\|^2$ • This is a PCA via SVD problem Gradient descent: • Minimum-norm sol. • $\beta^{(t+1)}=\beta^{(t)}-\eta\nabla LO_\beta(\beta^{(t)})=\beta^{(t)}-\eta 2(X^\top X\beta^{(t)}-X^\top y)$ • If X has orthonormal columns, SGD is initialized to $\beta^{(0)}=X^\top y$ i.e. $X\beta^{(0)}=XX^\top y$ from which we can see that, since X is orthonormal, $XX^\top y$ is the projection of y to the column space of X • Yields same result as OLSE

Weighted regression via orthogonality principle — • $\hat{h}=\arg\min_h\sum_{k=1}^m w_k|A_kh-x_k|^2=\sum_{k=1}^m|\sqrt{w_k}A_kh-\sqrt{w_k}x_k|^2=\sum_{k=1}^m|A'_kh-x'_k|^2=\|A'h-x'\|$ with $A'_k=\sqrt{w_k}A_k$ and $x'_k=\sqrt{w_k}x_k$ • Minimizes the weighted norm: $\|Ah-x\|_W^2=(Ah-x)^\top W(Ah-x)=\sum_{k=1}^m w_k|(Ah-x)_k|^2$ since W is diagonal $=\sum_{k=1}^m w_k|A_kh-x_k|^2$ • Satisfies orthogonality for weighted inner product: $\forall j:0=\langle Ah-x,A^{(j)}\rangle_W=A^{(j)\top}W(Ah-x)=A^{(j)\top}W^{1/2}W^{1/2}(Ah-x)=A^{(j)\top}(W^{1/2})^\top W^{1/2}(Ah-x)=A^{(j)\top}(W^{1/2})^\top(W^{1/2}Ah-W^{1/2}x)=(W^{1/2}A^{(j)})^\top(W^{1/2}Ah-W^{1/2}x)=(A^{(j)'})^\top(A'h-x')$

Hypothesis Testing of Found Parameters — • Let $y|X\sim N(y,\sigma^2I)=N(X\beta,\sigma^2I)$ • Let $\hat{\beta}=(X^\top X)^{-1}X^\top y=X^+y$ be the OLSE where X^+ is a scalar • Then, $\hat{\beta}\sim N(X^+X\beta,X^{+\top}\sigma^2X^+)=N(\beta,(X^\top X)^{-1}\sigma^2)$. Proof: ◦ $N(X^+X\beta,X^{+\top}\sigma^2X^+)=N(I\beta,\sigma^2X^+X^{+\top})$ since X^+ is a scalar ◦ Further, we have $N(I\beta,\sigma^2X^+((X^\top X)^{-1}X^\top)^\top)=N(\beta,\sigma^2X^+X((X^\top X)^{-1})^\top)=N(\beta,\sigma^2(X^\top X)^{-1})$ since $(X^\top X)$ is symmetric • We can est. σ^2 unbiasedly as: $\hat{\sigma}^2=\frac{1}{n-m}\sum_{i\le n}(X\hat{\beta}-y)^2$ • Then, confidence interval for $\hat{\beta}_j$ given by: $\hat{\beta}_j\pm z_{\alpha/2}\hat{se}(\hat{\beta}_j)$ where ◦ $z_{\alpha/2}=\Phi^{-1}(\alpha/2)$ is Gaussian CDF ◦ $\hat{se}(\hat{\beta}_j)$ is the j^{th} diagonal element of the covariance matrix $\sigma^2(X^\top X)^{-1}$ • We can perform a hypothesis test on $\hat{\beta}$ with the *Wald test*: ◦ $H_0:\beta=\beta_0$ (typically 0), $H_1:\beta\neq\beta_0$ ◦ Wald statistic: $W=\frac{\hat{\beta}-\beta_0}{\hat{se}}$ ◦ If p-value associated with W is smaller than α resp. if $|W|$ is greater than or equal to the critical value $z_{\alpha/2}$, we reject H_0

Evaluation — • OLSE is unbiased if noise ϵ has zero mean: ◦ Given $y=X\beta+\epsilon$, we can substitute $\hat{\beta}=(X^\top X)^{-1}X^\top(X\beta+\epsilon)=\beta+(X^\top X)^{-1}X^\top\epsilon$ ◦ Taking the expected value on both sides, we have: $\mathbb{E}[\hat{\beta}]=\beta+(X^\top X)^{-1}X^\top\mathbb{E}(\epsilon)$ ◦ Then, $\mathbb{E}(\hat{\beta})=\beta$ if the noise has zero mean • *Gauss Markov theorem*: OLSE is best (lowest variance, lowest MSE) unbiased est.

Characteristics — • Convex with psd Hessian • Has global min. • Has unique or infinitely many solutions • Can be solved analytically, if $X^\top X$ is invertible • In case of *multicollinearity*: ◦ The rank of X is less than full, i.e. there are multiple columns (predictor variables) that are linearly dependent ◦ $X^\top X$ is singular, i.e. non-invertible ◦ There are multiple solutions for β • If it has infinitely many solutions, the preferred sol. is the *min.-norm sol.*, which minimizes $\|\beta\|$ and lies in the column space of $X^\top$ resp. is a sol. to Xu

Linear Minimum Mean Squared Error Estimation (LMMSE)

Descr. • Min. MSE of two random variables, leveraging information about their mean and covariance • Linear regr. with large samples is a data-based proxy for LMMSE, since $\frac{1}{n}\mathbb{E}[X^\top X]=\mathbb{E}[yy^\top]$ and $\frac{1}{n}\mathbb{E}[X^\top y]=\mathbb{E}[xy]$ and as $n\to\infty$ the strong law of large numbers applies

Form. • $y=\alpha x+z$ is a vector of variables and is observed • α is a known vector • z a zero-mean noise vector • x is a random variable and quantity of interest • We est. x as $\hat{x}=h^\top y=\sum_i h_iy_i$ • This can be considered as a projection of x to the space spanned by y

Opt. *Parameters —* Find parameters h *Objective funct. —* • Min.: $LO=\mathbb{E}[(\hat{x}-x)^2]$ resp. $LO=\mathbb{E}[|\hat{x}-x|]$ *Opt. —* • By the orthogonality principle, $\mathbb{E}[(\hat{x}-x)y_i]=\mathbb{E}[(h^\top y-x)y_i]=\mathbb{E}[(\sum_{l=1}^n h_ly_l-x)y_i]=0$ for $i=1,\dots,n$ • Then, $h^\top\mathbb{E}[y_iy_l]=\sum_{l=1}^n\mathbb{E}[y_ly_i]h_l=\mathbb{E}[xy_i]$ for $i=1,\dots,n$ which in matrix notation corresponds to $\begin{bmatrix}\mathbb{E}[y_1y_1]&\dots&\mathbb{E}[y_1y_n]\\\vdots&\vdots&\vdots\\\mathbb{E}[y_ny_1]&\dots&\mathbb{E}[y_ny_n]\end{bmatrix}\begin{bmatrix}h_1\\\vdots\\h_n\end{bmatrix}=\begin{bmatrix}\mathbb{E}[xy_1]\\\vdots\\\mathbb{E}[xy_n]\end{bmatrix}$ resp. concisely $\mathbb{E}[yy^\top]h=\mathbb{E}[xy]$ where output is column vector (similar to linear regr.), or $h^\top\mathbb{E}[yy^\top]=\mathbb{E}[xy]$ where output is row vector (peculiar to LMMSE) • Then, $h^\top=\frac{\mathbb{E}[xy^\top]}{\mathbb{E}[yy^\top]}$ • If $y=x+z$ where x,z are independent and $\mathbb{E}(z)=0$: ◦ $\mathbb{E}[xy^\top]=\mathbb{E}[x(x+z)^\top]=\mathbb{E}[x^2]=\mathbb{V}[x]+\mathbb{E}[x]^2$ ◦ $\mathbb{E}[yy^\top]=\mathbb{E}[(x+z)(x+z)^\top]=\mathbb{E}[x^2]+\mathbb{E}[z^2]=\mathbb{V}[x]+\mathbb{E}[x]^2+\mathbb{E}[z^2]$ ◦ $h^\top=\frac{\mathbb{E}[x^2]}{\mathbb{E}[x^2]+\mathbb{E}[z^2]}=\frac{\mathbb{V}[x]+\mathbb{E}[x]^2}{\mathbb{V}[x]+\mathbb{E}[x]^2+\mathbb{E}[z^2]}$ • If $y=\alpha x+z$ where x,z are independent and $\mathbb{E}(z)=0$: ◦ $\mathbb{E}[xy^\top]=\mathbb{E}[x(\alpha x+z)^\top]=\alpha\mathbb{E}[x^2]=\alpha(\mathbb{V}[x]+\mathbb{E}[x]^2)$ ◦ $\mathbb{E}[yy^\top]=\mathbb{E}[(\alpha x+z)(\alpha x+z)^\top]=\alpha^2\mathbb{E}[x^2]+\mathbb{E}[z^2]=\alpha^2(\mathbb{V}[x]+\mathbb{E}[x]^2)+\mathbb{E}[z^2]$ ◦ $h^\top=\frac{\alpha\mathbb{E}[x^2]}{\alpha^2\mathbb{E}[x^2]+\mathbb{E}[z^2]}=\frac{\alpha(\mathbb{V}[x]+\mathbb{E}[x]^2)}{\alpha^2(\mathbb{V}[x]+\mathbb{E}[x]^2)+\mathbb{E}[z^2]}$ ◦ $\mathbb{E}[\hat{X}|X=x]=h^\top\mathbb{E}[y|X=x]=h^\top\alpha x$ • If $\hat{x}=\beta\alpha^\top y$ and $y=\alpha x+z$ where x,z are independent and $\mathbb{E}(z)=0$: ◦ To ensure unbiasedness: $\mathbb{E}[\hat{x}|x]=x$ ◦ We can plug in: $\mathbb{E}[\beta\alpha^\top(\alpha x+z)|x]=\beta\alpha^\top\alpha x=x\Rightarrow\beta=\frac{1}{\|\alpha\|^2}$ ◦ To minimize MSE: We can rewrite to $\tilde{x}=\beta\tilde{y}$ and $\tilde{y}=\alpha^\top(\alpha x+z)=\|\alpha\|^2x+\alpha^\top z$ ◦ $\mathbb{E}[x\tilde{y}]=\mathbb{E}[\tilde{y}]$ ◦ $\Rightarrow\beta=\frac{\mathbb{E}[x^2]}{\|\alpha\|^2\mathbb{E}[x^2]+\mathbb{E}[z^2]}$

Further proofs — • $\mathbb{E}[yy^\top]$ is not invertible if there exists an $a\neq0$ such that $y^\top a=0$ with probability 1. Proof: ◦ Both y and a are in $\mathbb{R}^n$ ◦ Direction 1: ■ If $y^\top a=0$, these two vectors are orthogonal. Then, y cannot span the entire space $\mathbb{R}^n$ ■ This means that some rows or columns in the matrix $yy^\top$ are linearly dependent ■ By the invertible matrix theorem, such matrices are not invertible ◦ Direction 2: ■ If the matrix $yy^\top$ is not invertible, by the invertible matrix theorem, it must have linearly dependent rows or columns ■ This means that one column or row can be expressed in terms of the sum over all other columns and rows ■ Then, the equation system $yy^\top a=0$ is underdetermined ■ Then, it is possible to formulate $y^\top a=0$ such that $a\neq0$

Bayesian Linear Regr.

Form. • $y^{(i)}=\beta\cdot x^{(i)}+\epsilon$ resp. $y=X\beta+\epsilon$ • $\beta\sim N(0,T^2I_m)$ • $p(\beta)\propto-\frac{1}{2T^2}\beta^\top\beta$ • $\epsilon\sim N(0,\sigma^2)$

Opt. *Parameters —* Find distribution of parameters β *Opt. —* • Prior $p(\beta)$: ◦ $\beta\sim N(0,T^2I_m)$ ◦ $p(\beta)=\frac{1}{(2\pi T^2)^{m/2}}exp(-\frac{1}{2T^2}\beta^\top\beta)$ • Likelihood $p(y|X,\beta)$: ◦ Conditional on β, $y\sim N(X\beta,\sigma^2I_n)$ ◦ $p(y|X,\beta)=\frac{1}{(2\pi\sigma^2)^{n/2}}exp(-\frac{1}{2\sigma^2}(y-X\beta)^\top(y-X\beta))$ • Posterior $p(\beta|X,y)$: ◦ $p(\beta|X,y)\propto p(y|X,\beta)\times p(\beta)\propto exp(-\frac{1}{2\sigma^2}(y-X\beta)^\top(y-X\beta))\times exp(-\frac{1}{2T^2}\beta^\top\beta)=exp(-\frac{1}{2}(\frac{1}{\sigma^2}y^\top y-2\beta^\top X^\top y+\beta^\top X^\top X\beta)+\frac{1}{2T^2}\beta^\top\beta)\propto exp(-\frac{1}{2}(\beta^\top(\frac{1}{\sigma^2}X^\top X+\frac{1}{2T^2}I_m)\beta)-\frac{2}{\sigma^2}\beta^\top X^\top y)$ ◦ We now apply a symmetric matrix property $x^\top Ax+2x^\top b=(x+A^{-1}b)^\top A(x+A^{-1}b)-b^\top A^{-1}b$, with $\beta=x$, $(\frac{1}{\sigma^2}X^\top X+\frac{1}{2T^2}I_m)=A$ and $(\frac{1}{\sigma^2}X^\top y)=b$ ◦ Through this, we get $p(\beta|X,y)\propto\exp(\frac{1}{2}(\beta+(\frac{1}{\sigma^2}X^\top X+\frac{1}{T^2}I_m)^{-1}(\frac{1}{\sigma^2}X^\top y))^\top(\frac{1}{\sigma^2}X^\top X+\frac{1}{T^2}I_m)(\beta+(\frac{1}{\sigma^2}X^\top X+\frac{1}{T^2}I_m)^{-1}(\frac{1}{\sigma^2}X^\top y)))$ ◦ Thus, $p(\beta|X,y)\sim N(\mu,\Sigma)$ with ■ $\mu=\Sigma\times\frac{1}{\sigma^2}X^\top y$ ■ $\Sigma=(\frac{1}{\sigma^2}X^\top X+\frac{1}{T^2}I_m)^{-1}$ • Conjugate prior

Ridge (ℓ_2) Regr.

Form. • $y^{(i)}=\beta\cdot x^{(i)}$ resp. $y=X\beta$

Opt. *Parameters —* Find parameters β subject to $\|\beta\|^2\le t$ resp. $\|\beta\|^2-t\le0$ *Objective funct. —* • Min. MSE subject to constr. • Lagr.: $LO=\frac{1}{n}\sum_{i=1}^n(y^{(i)}-\beta\cdot x^{(i)})^2+\lambda(\|\beta\|^2-t)$ resp. $LO=(y-X\beta)^\top(y-X\beta)+\lambda(\|\beta\|^2-t)$ *Opt. —* • $\nabla_\beta LO=0\Rightarrow\beta=(X^\top X+\lambda I)^{-1}X^\top y$ *Alternative formulations —* Still a OLSE problem: • We can rewrite the objective to min. $\|X\beta-y\|^2+\lambda\|\beta\|^2=\|\begin{bmatrix}X\beta-y\\\sqrt{\lambda}\beta\end{bmatrix}\|^2$ as the objective to min. $\|X'\beta-y'\|^2$ with $X'=\begin{bmatrix}X\\\sqrt{\lambda}I\end{bmatrix}$ and $y'=\begin{bmatrix}y\\0\end{bmatrix}$ Bayesian MAP: • Posterior $p(\beta|X,y)\propto$Likelihood $p(y|X,\beta)\sim N(X\beta,\sigma^2I_n)\times$ Prior $p(\beta)\sim N(0,\tau^2I_m)$: ◦ $p(\beta|X,y)\propto\log(exp(-\frac{1}{2\sigma^2}(y-X\beta)^\top(y-X\beta))\times exp(-\frac{1}{2T^2}\beta^\top\beta))\propto-\frac{\|y-X\beta\|^2}{\sigma^2}-\frac{\|\beta\|^2}{\tau^2}$ • Max. log post. (MAP): $\arg\min_\beta\frac{\|y-X\beta\|^2}{\sigma^2}+\frac{\|\beta\|^2}{\tau^2}=\|y-X\beta\|^2+\frac{\sigma^2}{\tau^2}\|\beta\|^2$ • Equiv. to log loss of regr. with $\lambda=\frac{\sigma^2}{\tau^2}$ Orthogonality principle: • Constraint $Ch=d$ • Let $\hat{h}$ and h be sol. of $Ch=d$, then: $C(h-\hat{h})=Ch-Ch=0$. • Then $h-\hat{h}$ is in the nullspace of C spanned by basis $B=[b_1,\dots,b_p]$ • Then we can express $h-\hat{h}$ in that basis: $h-\hat{h}=Bh'$ • From this we get: $h=Bh'+\hat{h}$ • Then we can formulate minimization problem: $\min_{h:Ch=d}\|Ah-x\|^2=\min_{h:h=Bh'+\hat{h}}\|Ah-x\|^2=\min_{h'}\|A(Bh'+\hat{h})-x\|^2=\|ABh'+A\hat{h}-x\|^2=\|A'h'-x'\|^2$ with $A'=AB$ and $x'=x-A\hat{h}$ *Effect —* • Shrinks certain elements of β to near 0. Proof: ◦ Gradient at optimality given by $\frac{\partial(y-X\beta)^\top(y-X\beta)}{\partial\beta}+2\lambda\beta=0$ ◦ Then, $\beta^*=-\frac{1}{2\lambda}\frac{\partial(y-X\beta)^\top(y-X\beta)}{\partial\beta}$ • Addresses multicollinearity: ◦ SVD for $X=USV^\top$ ◦ We can show that $X\beta=US(S^2+\lambda I)^{-1}SU^\top Y$. Proof: ■ $X\beta^r=X(X^\top X+\lambda I)^{-1}X^\top Y$ ■ $=UDV^\top((UDV^\top)^\top UDV^\top+\lambda I)^{-1}(UDV^\top)^\top Y$ ■ $=UDV^\top(VDU^\top UDV^\top+\lambda I)^{-1}VDU^\top Y$ ■ $=UDV^\top(VD^2V^\top+\lambda VV^\top)^{-1}VDU^\top Y$ ■ $=UDV^\top(V(D^2+\lambda I)V^\top)^{-1}VDU^\top Y$ ■ $=UDV^\top V(D^2+\lambda I)^{-1}V^\top VDU^\top Y$ ■ $=UD(D^2+\lambda I)^{-1}DU^\top Y$ ◦ In case of multicollinearity, the rank of X is less than full, and S^2 cannot be inverted. By adding λI to S, ridge regr. ensures that the equation remains solvable even if S is not invertible on its own

Characteristics — • Strictly with pd Hessian, since Lagr. term is strictly convex and the sum of a strictly convex funct. with a convex funct. is strictly convex • Has global min. • Has unique sol., as $(X^\top X+\lambda I)$ has linearly independent columns • Can be solved analytically, as $(X^\top X+\lambda I)$ is always invertible

Lasso (ℓ_1) Regr.

Form. • $y^{(i)}=\beta\cdot x^{(i)}$ resp. $y=X\beta$

Opt. *Parameters —* Find parameters β subject to $|\beta|\le t$ resp. $|\beta|-t\le0$ *Objective funct. —* • Min. MSE subject to constr. • Lagr.: $LO=\frac{1}{n}\sum_{i=1}^n(y^{(i)}-\beta\cdot x^{(i)})^2+\lambda(|\beta|-t)$ resp. $LO=(y-X\beta)^\top(y-X\beta)+\lambda(|\beta|-t)$ *Alternative formulations —* Bayesian MAP: • Posterior $p(\beta|X,y)\propto$Likelihood $p(y|X,\beta)\sim N(X\beta,\sigma^2I_n)\times$ Prior $p(\beta)\sim$Lapl.$(0,b)$: ◦ $p(\beta|X,y)\propto\log(\exp(-\frac{1}{2\sigma^2}(y-X\beta)^\top(y-X\beta))\times\prod_{j=1}^m\exp(-\frac{|\beta_j|}{b}))=\log p(\beta|X,y)\propto-\frac{\|y-X\beta\|^2}{\sigma^2}-\frac{\|\beta\|_1}{b}$ • Max. log post. (MAP): $\arg\min_\beta\frac{\|y-X\beta\|^2}{\sigma^2}+\frac{\|\beta\|_1}{b}=\|y-X\beta\|^2+\frac{\sigma^2}{b}\|\beta\|_1$ • Equiv. to log loss of regr. with $\lambda=\frac{\sigma^2}{b}$ *Effect —* • Shrinks certain elements of β to 0. Proof: ◦ Gradient at optimality given by $\frac{\partial(y-X\beta)^\top(y-X\beta)}{\partial\beta}+\frac{\partial\lambda|\beta|}{\partial\beta}=0$ ◦ $\frac{\partial\lambda|\beta|}{\partial\beta}$ non-differentiable because there is a sharp edge at $\beta=0$, but we can work with subgradients for $\beta\neq0$: -1 if $\beta<0$, 1 if $\beta>0$, else 0 ◦ If we have $-\lambda<\frac{\partial(y-X\beta)^\top(y-X\beta)}{\partial\beta}<\lambda$ the optimum is given by $\beta=0$ ◦ This means that some parameters are set to 0

Kernel Methods

Background on Kernel Methods *Kernel trick —* • If a prediction funct. is described solely in terms of inner products in the input space (not appl. to regularized regr.), it can be lifted into the feature space by replacing the inner product with the kernel funct. • $k(x^{(i)},x^{(j)})=\varphi(x^{(i)})\cdot\varphi(x^{(j)})\Leftrightarrow$ • Kernel matrix is psd $(\alpha^\top K\alpha=\|\sum_i\alpha^i\varphi(x^i)\|^2\ge0)$ and symmetric $(k(x^{(i)},x^{(j)})=k(x^{(j)},x^{(i)}))$ • Kernel trick requires that span of training instances $span(\varphi(x^{(1)}),\dots,\varphi(x^{(N)}))=\mathbb{R}^k$ and, thus, that $N\ge k$. Proof: $dim(span(\dots))=\begin{cases}N&\text{if }N<k\\k&\text{if }N\ge k\end{cases}$ • Allows to avoid directly seeking the k parameters, but only the n parameters that characterize α • Allows to avoid calculating $\varphi(z)$ when evaluating novel instance, but only sum over weighted set of n kernel funct. outputs

Form. • Feature map $\varphi:\mathbb{R}^m\to\mathbb{R}^k$ • Prediction funct.: $\beta\cdot\varphi(x^{(i)})$ • Cost funct.: $LO=\sum_{i=1}^n LO(y^{(i)},\beta\cdot\varphi(x^{(i)}))+\Omega(\beta))$ • Iff $\Omega(\beta)$ is a non-decreasing funct.: $\beta=\Phi^\top\alpha=\sum_{i=1}^n\alpha^i\varphi(x^i)$ • Pred.: $\beta\cdot\varphi(z)=\Phi^\top\alpha\cdot\varphi(z)=\sum_{i=1}^n\alpha^i\varphi(x^i)\cdot\varphi(z)=\sum_{i=1}^n\alpha^ik(x^i,z)$ • If the dim. of the feature space is k, we require $n\ge k$ training points so that $\varphi(x^1),\dots,\varphi(x^n)$ span $\mathbb{R}^k$ • Act of prediction becomes act of measuring similarity to training instances

Kernel Types *RBF kernel —* • $\varphi(x)=exp(-\frac{1}{2}\|x\|^2)[\frac{x^\alpha}{\sqrt{\alpha!}}]_{\alpha\in\mathbb{N}^m}$ • $k(x^{(i)},x^{(j)})=\sigma^2exp(-\frac{\|x^{(i)}-x^{(j)}\|^2}{2l^2})$. Proof: ◦ $exp(-\frac{1}{2}\|x^{(i)}\|^2)exp(-\frac{1}{2}\|x^{(j)}\|^2)\sum_\alpha[\frac{x^{(i)\alpha}x^{(j)\alpha}}{\alpha!}]$ ◦ Given multinomial series expansion, $\sum_\alpha[\frac{x^{(i)\alpha}x^{(j)\alpha}}{\alpha!}]=exp(x^{(i)\top}x^{(j)})$ ◦ Then, we get $exp(-\frac{1}{2}\|x^{(i)}\|^2-\frac{1}{2}\|x^{(j)}\|^2+x^{(i)\top}x^{(j)})=exp(-\frac{\|x^{(i)}-x^{(j)}\|^2}{2})$ • Gives access to infinite feature space • $l\nearrow$covariance over long distances$\nearrow$smoothness$\nearrow$ • $\sigma\nearrow$vertical scale$\nearrow$ • *Stationary*, meaning that only the relative distance between two points determines the value output by the kernel funct. • Challenge: Cannot ignore irrelevant dimensions (whereas e.g. a neural network can do this by setting the associated weights to 0) *Periodic kernel —* • $k(x^{(i)},x^{(j)})=\sigma^2exp(-\frac{2\sin^2\frac{\pi\|x^{(i)}-x^{(j)}\|}{p}}{l^2})$ • $p\nearrow$period length$\nearrow$ *Kernel compositions —* • $k'(x_1,x_2)=ck(x_1,x_2)$, since $\varphi'(x)=\sqrt{c}\varphi(x)$ • $k'(x_1,x_2)=f(x_1)k(x_1,x_2)f(x_2)$, since $\varphi'(x)=f(x)\varphi(x)$ • $k'(x_1,x_2)=k_1(x_1,x_2)+k_2(x_1,x_2)$, since $\varphi'(x)=[\varphi_1(x)\quad\varphi_2(x)]^\top$ • $k'(x_1,x_2)=k_1(x_1,x_2)k_2(x_1,x_2)$, since $\varphi'(x)=\varphi_1(x)\otimes\varphi_2(x)$ (elw. tensor prod.) • $k'(x_1,x_2)=k(\psi(x_1),\psi(x_2)))$, since $\varphi'(x)=\varphi(\psi(x))$ • $k'(x_1,x_2)=p(k(x_1,x_2))$ where p is a polyn., since $=\sum_i a_ik_i(x_1,x_2)$ cf. above • $k'(x_1,x_2)=\exp(k(x_1,x_2))$, since acc. to Taylor exp. $\sum_{n=1}^r\frac{k(x_1,x_2)^r}{r!}=exp(k(x_1,x_2))=k'(x_1,x_2)$ cf. above • $k'(x_1,x_2)=x_1^\top Ax_2$ for psd, symm. $A=L^\top L$ acc. to Chol. decomp.

Kernel Ridge Regr.

Form. • $y=\beta\cdot\varphi(x^{(i)})$

Opt. *Objective funct. —* • Min. MSE: $LO=\frac{1}{n}\sum_{i=1}^n(y^{(i)}-\beta\cdot\varphi(x^{(i)}))^2$ *Opt. —* • Primal solution: ◦ Parameters: $\beta=(\Phi^\top\Phi+\lambda I)^{-1}\Phi^\top y$ ◦ Pred.: $\beta\cdot\varphi(z)=(\Phi^\top\Phi+\lambda I)^{-1}\Phi^\top y\cdot\varphi(z)=y^\top\Phi(\Phi^\top\Phi+\lambda I)^{-1}\varphi(z)$ • Define $K=\Phi\Phi^\top$ with $K_{ij}=\varphi(x^i)\cdot\varphi(x^{(j)})$ • Dual solution: ◦ Parameters: $\beta=\Phi^\top\alpha=\Phi^\top(K+\lambda I)^{-1}y$ Proof 1: ■ $(\Phi^\top\Phi+\lambda I_D)\beta=\Phi^\top y$ ■ $\Rightarrow\Phi^\top\Phi\beta+\lambda I_D\beta=\Phi^\top y$ ■ $\Rightarrow I_D\beta=\Phi^\top\lambda^{-1}(y-\Phi\beta)$ ■ Since we know from the representer theorem that $\beta=\Phi^\top\alpha$, we can say: $\alpha=\lambda^{-1}(y-\Phi\beta)$ ■ We can further develop this to: $\lambda\alpha=(y-\Phi\beta)$ ■ Replacing β by $\Phi^\top\alpha$ yields: $\lambda\alpha=(y-\Phi\Phi^\top\alpha)$ ■ $\Rightarrow\alpha=(\Phi\Phi^\top+\lambda I_N)^{-1}y=K^{-1}y$ ■ With this, we can calculate the parameters: $\beta=\Phi^\top\alpha=\Phi^\top(\Phi\Phi^\top+\lambda I_N)^{-1}y=\Phi^\top K^{-1}y$ Proof 2: ■ Theorem: $(FH^{-1}G+E)^{-1}FH^{-1}=E^{-1}F(GE^{-1}F+H)^{-1}$ ■ With $E=I_D,F=\Phi^\top,G=\Phi,H=I_N$, simplify to $(\Phi^\top\Phi+I_D)^{-1}\Phi^\top=I_D^{-1}\Phi^\top(\Phi\Phi^\top+I_N)^{-1}$ ■ Since $I_D^{-1}=I_D$, we have: $(\Phi^\top\Phi+I_D)^{-1}\Phi^\top=\Phi^\top(\Phi\Phi^\top+I_N)^{-1}$ ■ $\Rightarrow(\Phi^\top\Phi+\lambda I_D)^{-1}\Phi^\top y=\Phi^\top(\Phi\Phi^\top+I_N)^{-1}y=\sum_i\varphi(x^i)[(\Phi\Phi^\top+I_N)^{-1}y]_i$ ◦ Pred.: $\beta\cdot\varphi(z)=\Phi^\top\alpha\cdot\varphi(z)=y^\top(\Phi\Phi^\top)^{-1}\Phi\varphi(z)=y^\top(\Phi\Phi^\top)^{-1}k$ where $k=\Phi\varphi(z)=[k(x^{(1)},z),\dots,k(x^{(n)},z)]^\top=[\varphi(x^{(1)})\cdot\varphi(z),\dots,\varphi(x^{(n)})\cdot\varphi(z)]^\top$

Kernel K-Means Clustering

Form. • We specify that we want $j=1,\dots,k$ clusters in total • Clusters defined by centroid $\mu^{[j]}$ • Instances $i=1,\dots,n$ given by $x^{(i)}$ • Clusters and instances can be lifted from input to feature space via mapping $\phi(\cdot)$ • Each instance i has k indicator variables, which describe whether instance i is assigned to cluster j, given by $\{p^{i[j]}\}_{j=1}^k\in[0,1]$

Opt. *Parameters —* Find centroids $\mu^{[j]}$, i.e., find cluster assignments (instance always assigned to closest cluster in feature space) *Objective funct. —* • Min. the distance between each instance and the centroid of its closest cluster in feature space: $j^*=\arg\min_j\|\phi(x^{(i)})-\phi(\mu^{[j]})\|^2$ • Distortion funct. given by: $\Theta=\sum_{i=1}^n\sum_{j=1}^k p^{i[j]}\|\phi(x^{(i)})-\phi(\mu^{[j]})\|^2=\sum_{x^{(i)}\in C_j}\sum_{j=1}^k\|\phi(x^{(i)})-\phi(\mu^{[j]})\|^2$ with $p^{i[j]}=\begin{cases}1&\text{if }j=j^*\\0&\text{otherwise}\end{cases}$ and $C_j=\{x^{(i)}|p^{i[j]}=1\}$ • $\phi(\mu^{[j]})$ is defined by mean of points in cluster C_j in feature space: $\frac{1}{|C_j|}\sum_{x^{(\mu)}\in C_j}\phi(x^{(\mu)})$ • Then, we can rewrite the distance explicitly in terms of the kernel funct.: $\Theta=\sum_{j=1}^k\sum_{x^{(i)}\in C_j}\Big|K(x^{(i)},x^{(i)})-\frac{2}{|C_j|}\sum_{x^{(\mu)}\in C_j}K(x^{(i)},x^{(\mu)})+\frac{1}{|C_j|^2}\sum_{x^{(\mu)},x^{(\rho)}\in C_j}K(x^{(\mu)},x^{(\rho)})\Big|$

Logistic Regr.

Form. • Probability of each class is estimated via sigmoid funct.: $\sigma(z)=\frac{1}{1+e^{-z}}=\frac{e^z}{1+e^z}$ • $P(y=1|x)=\frac{1}{1+e^{-\beta\cdot x}}=\frac{e^{\beta\cdot x}}{1+e^{\beta\cdot x}}$ • $P(y=0|x)=\frac{1}{1+e^{\beta\cdot x}}=\frac{e^{-\beta\cdot x}}{1+e^{-\beta\cdot x}}$ • Odds: $\frac{P(y=1|x)}{P(y=0|x)}=e^{\beta\cdot x}$ • Log-odds: $ln(\frac{P(y=1|x)}{P(y=0|x)})=\beta\cdot x=ln(\frac{P(y=1)}{P(y=0)})+ln(\frac{P(x|y=1)}{P(x|y=0)})$, i.e. can be decomposed into prior and likelihood • We can influence the prior by introducing an offset to the log-odds: $\beta\cdot x+t$: ◦ In multinomial case: $p_k=\frac{e^{z_k+t_k}}{\sum_{\ell=1}^k e^{z_\ell+t_\ell}}=\frac{e^{z_k}e^{t_k}}{\sum_{\ell=1}^k e^{z_\ell}e^{t_\ell}}=\frac{p(x_k,y)e^{t_k}}{\sum_{\ell=1}^k p(x_\ell,y)e^{t_\ell}}$ where $t_k=\ln(\frac{q(x_k)}{p(x_k)})$ ◦ Then, the prior on x ($p(x)$) is replaced by $q(x)$ • Geometrically, $z=\beta\cdot x$ defines a linear separating hyperplane: ◦ When $z>0$ resp. log-odds >0, then odds >1 resp. $P(y=1|x)>P(y=0|x)$ resp. $\sigma(z)>0.5$, then predict $y=1$ ◦ When $z<0$ resp. log-odds <0, then odds <1 resp. $P(y=1|x)<P(y=0|x)$ resp. $\sigma(z)<0.5$, predict $y=0$ ◦ When $z=0$ resp. log-odds =0, then odds =1 resp. $P(y=1|x)=P(y=0|x)$ resp. $\sigma(z)=0.5$, then we are on the decision boundary ◦ Decision boundary z is orthogonal to β. Proof: ■ For any 2 points x_1,x_2 on the decision boundary $z=0$, i.e. $\beta\cdot x_1=0$, $\beta\cdot x_2=0$ ■ Since x_1,x_2 are on the decision boundary, vector z can be considered a linear combination of x_1-x_2 ■ Combining these equations, we get $\beta\cdot(x_1-x_2)=\beta\cdot uz=0$

Opt. *Objective funct. —* • Likelihood: $L(\beta)=\prod_{i=1}^n P(y^{(i)}|x^{(i)};\beta)=\prod_{i=1}^n\sigma(z^{(i)})^{y^{(i)}}(1-\sigma(z^{(i)}))^{1-y^{(i)}}$ • Max. log-likelihood: $\log L(\beta)=\sum_{i=1}^n[y^{(i)}\log\sigma(z^{(i)})+(1-y^{(i)})\log(1-\sigma(z^{(i)}))]=\sum_{i=1}^n[y^{(i)}\log\frac{1}{1+e^{-z^{(i)}}}+(1-y^{(i)})\log\frac{e^{-z^{(i)}}}{1+e^{-z^{(i)}}}]=\sum_{i=1}^n[y^{(i)}z^{(i)}-\log(1+e^{z^{(i)}})]$ • Min. log-loss: $-\log L(\beta)$ *Opt. —* • $\frac{\partial-\log L(\beta)}{\partial\beta}=-\sum_{i=1}^n\frac{\partial}{\partial\beta}[y^{(i)}\log\sigma(z^{(i)})+(1-y^{(i)})\log(1-\sigma(z^{(i)}))]=\sum_{i=1}^n[\sigma(z^{(i)})-y^{(i)}]\frac{\partial z^{(i)}}{\partial\beta}=\sum_{i=1}^n[\sigma(z^{(i)})-y^{(i)}]x^{(i)}$. Proof: ◦ Derivative of the sigmoid: $\frac{\partial\sigma(z^{(i)})}{\partial z^{(i)}}=\sigma(z^{(i)})(1-\sigma(z^{(i)}))$ ◦ Derivative of the first term: $\frac{\partial}{\partial\beta}[y^{(i)}\log\sigma(z^{(i)})]=y^{(i)}\frac{1}{\sigma(z^{(i)})}\frac{\partial\sigma(z^{(i)})}{\partial z^{(i)}}\frac{\partial z^{(i)}}{\partial\beta}=y^{(i)}\frac{1}{\sigma(z^{(i)})}\sigma(z^{(i)})(1-\sigma(z^{(i)}))x^{(i)}=y^{(i)}(1-\sigma(z^{(i)}))x^{(i)}=y^{(i)}x^{(i)}-y^{(i)}\sigma(z^{(i)})x^{(i)}$ ◦ Derivative of the second term: $\frac{\partial}{\partial\beta}[(1-y^{(i)})\log(1-\sigma(z^{(i)}))]=(1-y^{(i)})\frac{1}{1-\sigma(z^{(i)})}(-1)\frac{\partial\sigma(z^{(i)})}{\partial z^{(i)}}\frac{\partial z^{(i)}}{\partial\beta}=(1-y^{(i)})\frac{1}{1-\sigma(z^{(i)})}(-1)\sigma(z^{(i)})(1-\sigma(z^{(i)}))x^{(i)}=-(1-y^{(i)})\sigma(z^{(i)})x^{(i)}=y^{(i)}\sigma(z^{(i)})x^{(i)}-\sigma(z^{(i)})x^{(i)}$

Convexity of Log Loss — • Sum of convex functions is convex • Thus, we need to prove convexity of $\ln(1+e^{\beta\cdot x^{(i)}})$ and $-y^{(i)}\beta\cdot x^{(i)}$ • For second term: ◦ $\mathcal{H}(\beta)=\nabla_\beta^2(-y^{(i)}\beta\cdot x^{(i)})=0\Rightarrow\mathcal{H}\succcurlyeq0$ • For first term: ◦ $\mathbb{H}(\beta)=\nabla_\beta^2\ln(1+e^{\beta\cdot x^{(i)}})$ ◦ $=v(x)\times u'(x)-v'(x)\times u(x)=\frac{1}{1+e^{\beta\cdot x}}\times e^{\beta\cdot x}\times xx^\top-\frac{e^{\beta\cdot x}x}{(1+e^{\beta\cdot x})^2}\times x^\top\times e^{\beta\cdot x}$ ◦ $=\frac{e^{\beta\cdot x}xx^\top(1+e^{\beta\cdot x})-e^{\beta\cdot x}xx^\top e^{\beta\cdot x}}{(1+e^{\beta\cdot x})^2}$ ◦ $=\frac{e^{\beta\cdot x}xx^T}{(1+e^{\beta\cdot x})^2}$ ◦ $\mathcal{H}\succ0$. Proof: $a^T\mathcal{H}a=\frac{e^{\beta\cdot x}}{(1+e^{\beta\cdot x})^2}a^Txx^Ta=\frac{e^{\beta\cdot x}}{(1+e^{\beta\cdot x})^2}\|a^Tx\|^2\ge0$

Regularization — • Perfectly separable data requires regularization • Let weights for each class k be scaled by c as $c\tilde{\beta}_k$ • Gradient of log-loss with respect to c is always negative, causing gradient descent to grow c without bound. Proof: ◦ Log loss$=\sum_{i=1}^n\ln(1+e^{c\tilde{\beta}\cdot x^{(i)}})-y^{(i)}c\tilde{\beta}\cdot x^{(i)}$ ◦ ∇_c log loss$=\sum_{i=1}^n\frac{1}{1+e^{c\tilde{\beta}\cdot x^{(i)}}}\times e^{c\tilde{\beta}\cdot x^{(i)}}\times\tilde{\beta}\cdot x^{(i)}-y^{(i)}\tilde{\beta}\cdot x^{(i)}$ ◦ $=\sum_{i=1}^n\tilde{\beta}\cdot x^{(i)}(\frac{e^{c\tilde{\beta}\cdot x^{(i)}}}{1+e^{c\tilde{\beta}\cdot x^{(i)}}}-y^{(i)})$ ◦ Given perfect separation: ■ If $y^{(i)}=1$, $\tilde{\beta}\cdot x^{(i)}>0$, $\frac{e^{c\tilde{\beta}\cdot x^{(i)}}}{1+e^{c\tilde{\beta}\cdot x^{(i)}}}-y^{(i)}<0$ ■ If $y^{(i)}=0$, $\tilde{\beta}\cdot x^{(i)}<0$, $\frac{e^{c\tilde{\beta}\cdot x^{(i)}}}{1+e^{c\tilde{\beta}\cdot x^{(i)}}}-y^{(i)}>0$ ◦ Thus, for all i, ∇_c log loss<0 ◦ Thus gradient descent will cause c to grow without bound

Multinomial Logistic Regr.

Form. • Probability of each class is estimated via the softmax funct. (generalizes the sigmoid funct. to multiple classes): $P(y=k|x)=\frac{e^{f_k(x)/T}}{\sum_{j=1}^k e^{f_j(x)/T}}=\frac{e^{\beta_k\cdot x/T}}{\sum_{j=1}^k e^{\beta_j\cdot x/T}}$ where T is the temperature and allows to smoothly interpolate between the differentiable softmax (T=1) and the non-differentiable argmax (T=0) • The predicted class is the one with the highest probability: $\hat{y}=\arg\max_k P(y=k|x)$ • Geometrically, the softmax defines $k-1$ linear separating hyperplanes

Opt. *Parameters* — Find parameters $\beta_1,\dots,\beta_k$

Objective funct. — • Likelihood: $L(\beta)=\prod_{i=1}^n P(y^{(i)}|x^{(i)};\beta)=\prod_{i=1}^n \prod_{\ell=1}^k \left(\frac{e^{\beta_\ell \cdot x^{(i)}}}{\sum_{j=1}^k e^{\beta_j\cdot x^{(i)}}}\right)^{\delta\{y^{(i)}=\ell\}}$ • Max. log-likelihood: $\log L(\beta)=\sum_{i=1}^n\sum_{\ell=1}^k \delta\{y^{(i)}=\ell\}[\beta_\ell\cdot x^{(i)} - \log(\sum_{j=1}^k e^{\beta_j\cdot x^{(i)}})]$ • Min. log-loss: $-\log L(\beta)$

Opt. — • $\frac{\partial -\log L(\beta)}{\partial \beta_k} = -\sum_{i=1}^n\sum_{\ell=1}^k \frac{\partial}{\partial\beta_k}[\delta\{y^{(i)}=\ell\}[\beta_\ell\cdot x^{(i)}-\log(\sum_{j=1}^k e^{\beta_j\cdot x^{(i)}})]]=-\sum_{i=1}^n \delta\{y^{(i)}=k\}x^{(i)} - P(y=k|x^{(i)})x^{(i)}$. Proof: ○ Derivative of the first term: $\frac{\partial}{\partial\beta_k}(\sum_{\ell=1}^k \delta\{y^{(i)}=\ell\}\beta_\ell\cdot x^{(i)})=\delta\{y^{(i)}=k\}x^{(i)}$ ○ Derivative of the second term: $\frac{\partial}{\partial\beta_k}(-\log(\sum_{j=1}^k e^{\beta_j\cdot x^{(i)}}))= -\frac{1}{\sum_{j=1}^k e^{\beta_j\cdot x^{(i)}}}\frac{\partial}{\partial\beta_k}(\sum_{j=1}^k e^{\beta_j\cdot x^{(i)}})= -\frac{e^{\beta_k\cdot x^{(i)}}}{\sum_{j=1}^k e^{\beta_j\cdot x^{(i)}}}x^{(i)}=-P(y=k|x^{(i)})x^{(i)}$

Decision Tree for Regr.

Form. • Predictor space is split along $J-1$ internal nodes • Thus, we get J distinct regions resp. terminal nodes $R_1,\dots,R_J$

Algorithm • Splitting approach: Recursive binary splitting (CART): ○ Top-down: Successively splits the predictor space ○ Greedy: At each step, the best split is made ○ Steps: 1) Select predictor x_j and threshold s such that splitting the predictor space into the regions $\{x_j\le s\}$ and $\{x_j>s\}$ leads to the smallest MSE 2) Repeat the process for one of the previously identified regions • Pred.: ○ For each observation falling into R_j, the prediction is the mean or median of the response values for the training observations in R_j, denoted as $\hat{y}_{R_j}$ ○ At each leaf node, 2 numbers are stored: The sum of outputs and the number of samples

Opt. *Parameters* — Find regions $R_1,\dots,R_J$

Objective funct. — • Min. $\sum_{j=1}^J\sum_{i\in R_j}(y_i-\hat{y}R_j)^2$

Decision Tree for Classification

Form. • Predictor space is split along $J-1$ internal nodes • Thus, we get J distinct regions resp. terminal nodes $R_1,\dots,R_J$

Algorithm • Splitting approach: Recursive binary splitting (CART): ○ Top-down: Successively splits the predictor space ○ Greedy: At each step, the best split is made ○ Steps: 1) Select predictor x_j and threshold s such that splitting the predictor space into the regions $\{x_j\le s\}$ and $\{x_j>s\}$ leads to the smallest Gini-index or cross-entropy 2) Repeat the process for one of the previously identified regions • Pred.: ○ For each observation falling into R_j, the prediction is the most common response class for the training observations in R_j, denoted as $\hat{y}_{R_j}$ ○ At each leaf node, k numbers are stored: Number of samples in each of the k classes

Opt. *Parameters* — Find regions $R_1,\dots,R_J$

Objective funct. — • Gini-index: $G_j=\sum_{k=1}^K \hat{p}_{jk}(1-\hat{p}_{jk})$ where $\hat{p}_{jk}$ is the proportion of training observations in R_j that are from class k ○ G_j takes a small value when nodes are pure, i.e., $\hat{p}_{jk}$ is 0 or 1 • Cross-entropy: $D_j=-\sum_{k=1}^K \hat{p}_{jk}\log(\hat{p}_{jk})$ where $\hat{p}_{jk}$ is the proportion of training observations in R_j ○ Alternatively $J(B)=-\frac{1}{h}\sum_{i=1}^h\sum_{k=1}^K y_i^{(k)}\log(\hat{p}_i^{(k)})$ where $y_i^{(k)}$ indicates whether the i^{th} training observation belongs to class k ○ D_j (or $J(B)$) takes a small value when nodes are pure, i.e., $\hat{p}_{jk}$ (or $\hat{p}_i^{(k)}$) is 0 or 1 • Information gain: Reduction in cross-entropy

K-Means Clustering

Form. • We specify that we want $j=1,\dots,k$ clusters in total • Clusters defined by centroid $\mu^{[j]}\in\mathbb{R}^m$ • Instances $i=1,\dots,n$ given by $x^{(i)}$ • Each instance i has k indicator variables, which describe whether instance i is assigned to cluster j, given by $\{p^{i[j]}\}_{j=1}^k\in[0,1]$

Opt. *Parameters* — Find centroids $\mu^{[j]}$, i.e. find cluster assignments (instance always assigned to closest cluster)

Objective funct. — • Min. distance between each instance and the centroid of its closest cluster: $j^*=\arg\min_j \sum_i\sum_j p^{i[j]}\|x^i-\mu^{[j]}\|^2$ • Distortion funct. given by: $\Theta=\sum_{i=1}^n\sum_{j=1}^k p^{i[j]}\|x^{(i)}-\mu^{[j]}\|^2=\sum_{x^{(i)}\in C_j}\sum_{j=1}^k\|x^{(i)}-\mu^{[j]}\|^2$ with $p^{i[j]}=\begin{cases}1 & \text{if } j=j^*\\0&\text{otherwise}\end{cases}$ and $C_j=\{x^{(i)}|p^{i[j]}=1\}$ • If we choose Euclidean distance: $\|x^{(i)}-\mu^{[j]}\|=\sqrt{\sum_{l=1}^m(x_l^{(i)}-\mu_l^{[j]})^2}$

Opt. — Lloyd's algorithm: 1) Randomly initialize each $\mu^{[j]}$ 2) E-Step: • Re-assign instances while keeping all centroids fixed, i.e. min. Θ with respect to $p^{i[j]}$ 3) M-Step: • Re-compute centroids while keeping all assignments fixed, i.e. min. Θ with respect to $\mu^{[j]}$ • $\nabla_{\mu^{[j]}}\Theta=2\sum_{i=1}^n p^{i[j]}(x^{(i)}-\mu^{[j]})=0$ • Then, $\mu^{[j]}=\frac{\sum_{i=1}^n p^{i[j]}x^{(i)}}{\sum_{i=1}^n p^{i[j]}}=\frac{1}{|C_j|}\sum_{x^{(i)}\in C_j}x^{(i)}$ • Θ is strictly convex, thus, we find a global min. and a unique sol. here 4) Repeat E- and M-Step until convergence Proof: • Let assignment funct. $c(i)\to j$ be defined as: $c(i)=\arg\min_{j\in\{1,\dots,k\}}\|x^{(i)}-\mu^{[j]}\|^2$ • Distortion can be rewritten as $\Theta=\sum_i\|x^{(i)}-\mu^{c(i)}\|^2$ • In E-Step, the centroids are fixed and we recalculate the assignment funct. $c(i)$. Thus, in E-Step Θ decreases unless all assignments $c(i)$ remain unchanged • In M-Step, the clusters C_j are fixed and we recalculate $\mu^{[j]}$ to min.: $\sum_{x^{(i)}\in C_j}\|x^{(i)}-\mu^{[j]}\|^2$ • $\mu^{[j]}=\frac{1}{|C_j|}\sum_{x^{(i)}\in C_j}x^{(i)}$ minimizes this cost: ○ For any vector $\mu^{[j]'}$, we can write the cost as: $\frac{1}{|C_j|}\sum_{x^{(i)}\in C_j}\|x^{(i)}-\mu^{[j]'}\|^2=\frac{1}{|C_j|}\sum_{x^{(i)}\in C_j}\|(x^{(i)}-\mu^{[j]})+(\mu^{[j]}-\mu^{[j]'})\|^2=\frac{1}{|C_j|}\sum_{x^{(i)}\in C_j}\|x^{(i)}-\mu^{[j]}\|^2+\|\mu^{[j]}-\mu^{[j]'}\|^2+2(\mu^{[j]}-\mu^{[j]'})^\top\frac{1}{|C_j|}\sum_{x^{(i)}\in C_j}(x^{(i)}-\mu^{[j]})$ ○ The third term vanishes since $\frac{1}{|C_j|}\sum_{x^{(i)}\in C_j}x^{(i)}=\mu^{[j]}$ ○ Then, we have: $=\frac{1}{|C_j|}\sum_{x^{(i)}\in C_j}\|x^{(i)}-\mu^{[j]}\|^2+\|\mu^{[j]}-\mu^{[j]'}\|^2$ ○ We can see that: $\frac{1}{|C_j|}\sum_{x^{(i)}\in C_j}\|x^{(i)}-\mu^{[j]}\|^2+\|\mu^{[j]}-\mu^{[j]'}\|^2\ge\frac{1}{|C_j|}\sum_{x^{(i)}\in C_j}\|x^{(i)}-\mu^{[j]}\|^2$ with eq. if $\mu^{[j]'}=\mu^{[j]}$

Principal Component Analysis (PCA)

Overview — • Axes = *principal components* • Axis = eigenvector with *loadings*, indicating how much each feature in the original data contributes • Variance along each axis = eigenvalue

Max. Variance Approach

Form. • Project data $\{x^{(i)}\}_{i=1}^n\in\mathbb{R}^m$ onto space $\mathbb{R}^d$ spanned by orthonormal basis $\{u^{[j]}\}_{j=1}^d\in\mathbb{R}^m$ where $d<<m$ • Each instance $x^{(i)}$ is projected onto each basis vectors $u^{[j]}\cdot x^{(i)}$: $x^{(i)}\to[u^{[1]}\cdot x^{(i)},\dots,u^{[d]}\cdot x^{(i)}]^\top$ • Each basis vector $u^{[j]}$ contains m loadings $[u_1^{[j]},\dots,u_m^{[j]}]$, whose value indicates how important each feature m is for the j^{th} principal component • Mean of projected data for a given basis vector: $u^{[j]}\cdot\bar{x}=u^{[j]}\cdot\frac{1}{n}\sum_{i=1}^n x^{(i)}$ • Variance of projected data for a given basis vector: $\frac{1}{n}\sum_{i=1}^n(u^{[j]}\cdot x^{(i)}-u^{[j]}\cdot\bar{x})^2=u^{[j]\top}Su^{[j]}$ where $S=\frac{1}{n}\sum_{i=1}^n(x^{(i)}-\bar{x})(x^{(i)}-\bar{x})^\top=\frac{1}{n}X^\top X$ is the covariance matrix of the data in the orthogonal complement to the subspace spanned by the first $j-1$ principal components, i.e. $X_{j-1}=\{x^{(i)}-\text{projection}_{u_{\le j-1}}\}=\{x^{(i)}-\sum_{l=1}^{j-1}(u^{[l]}\cdot x^{(i)})u^{[l]}\}$

Opt. *Parameters* — Find principal components $\{u^{[j]}\}_{j=1}^d$

Objective funct. — • For the j^{th} principal component, max. variance $\sum_{j=1}^d u^{[j]\top}Su^{[j]}$ subject to orthonormal $\{u^{[j]}\}_{j=1}^d$ • Gives rise to Lagr. • Lagr. for $u^{[1]}$ capturing the most variance: $\mathcal{L}=u^{[1]\top}Su^{[1]}-\lambda^{[1]}(u^{[1]}\cdot u^{[1]}-1)$ where $\lambda^{[1]}$ captures the orthonormality constr. that $u^{[1]}\cdot u^{[1]}=1$ • Lagr. for $u^{[2]}$ capturing the second-most variance: $\mathcal{L}=u^{[2]\top}Su^{[2]}-\lambda^{[2]}(u^{[2]}\cdot u^{[2]}-1)-\lambda^{[1][2]}(u^{[1]}\cdot u^{[2]}-0)$ with addtl. orthogonality constr.

Opt. — For $u^{[1]}$, based on $X=X_0$: • $\nabla_{u^{[1]}}\mathcal{L}=2Su^{[1]}-2\lambda^{[1]}u^{[1]}=0$ • $\Rightarrow Su^{[1]}=\lambda^{[1]}u^{[1]}$ • This is the eigenvector/eigenvalue equation, so $u^{[1]}$ is the eigenvector of S associated with the largest eigenvalue $\lambda^{[1]}$ • We see that the variance of the projected data is equal to $\lambda^{[1]}$: $u^{[1]\top}Su^{[1]}=u^{[1]\top}\lambda^{[1]}u^{[1]}=\lambda^{[1]}u^{[1]\top}u^{[1]}=\lambda^{[1]}\times1$ In the end, we have a total projected variance of $\sum_{j=1}^d\lambda^{[j]}$

SVD Approach

Form. • Project data $\{x^{(i)}\}_{i=1}^n\in\mathbb{R}^m$ onto space $\mathbb{R}^d$ spanned by orthonormal basis $\{u^{[j]}\}_{j=1}^d\in\mathbb{R}^m$ (subset of $\{u^{[j]}\}_{j=1}^m\in\mathbb{R}^m$) where $d\ll m$ • Given basis, we can represent original datapoint as: $x^{(i)}=\sum_{j=1}^m(x^{(i)}\cdot u^{[j]})u^{[j]}$ • We can represent projected datapoint as: $\tilde{x}^{(i)}=BB^\top x^{(i)}=B^\top x^{(i)}$ since B is orthogonal or, equivalently $\tilde{x}^{(i)}=\sum_{j=1}^d\alpha_{ij}u^{[j]}+\sum_{j=d+1}^m\gamma_j u^{[j]}$ where α_{ij} is specific to the instance and γ_j is generic and maps up to the subspace

Opt. *Objective funct.* — • Reconstruction error: $J=\frac{1}{n}\sum_{i=1}^n\|x^{(i)}-\tilde{x}^{(i)}\|^2=\frac{1}{n}\sum_{i=1}^n\|x^{(i)}\|^2-\|\tilde{x}^{(i)}\|^2=\frac{1}{n}\sum_{i=1}^n\|x^{(i)}\|^2-\|B^\top x^{(i)}\|^2$ • We can show that $\alpha_{ij}=x^{(i)}\cdot u^{[j]}$ and $\gamma_j=\bar{x}\cdot u^{[j]}$. Proof: ○ Take derivative of reconstruction error J and set it to 0 • If $m=d-1$, $J=u^{[d]\top}Su^{[d]}$. Proof: ○ For $d=m-1$, $\tilde{x}^{(i)}=\sum_{j=1}^d\alpha_{ij}u^{[j]}$ ○ Substituting $\tilde{x}^{(i)}$ and $x^{(i)}$ in J, we get: $J=\frac{1}{n}\sum_{i=1}^n\|x^{(i)}-\tilde{x}^{(i)}\|^2=\frac{1}{n}\sum_{i=1}^n\|\sum_{j=1}^m(x^{(i)}\cdot u^{[j]})u^{[j]}-\sum_{j=1}^{m-1}\alpha_{ij}u^{[j]}\|^2=\frac{1}{n}\sum_{i=1}^n\|\sum_{j=1}^m(x^{(i)}\cdot u^{[j]})u^{[j]}-\sum_{j=1}^{m-1}(x^{(i)}\cdot u^{[j]})u^{[j]}\|^2=\frac{1}{n}\sum_{i=1}^n\|(x^{(i)}\cdot u^{[D]})u^{[D]}\|^2$ ○ Using the orthonormality of $u^{[D]}$, we simplify: $\|(x^{(i)}\cdot u^{[D]})u^{[D]}\|^2=(x^{(i)}\cdot u^{[D]})^2$ ○ Thus: $J=\frac{1}{n}\sum_{i=1}^n(x^{(i)}\cdot u^{[D]})^2$ ○ To relate this to the covariance matrix $S=\frac{1}{n}\sum_{i=1}^n(x^{(i)}-\bar{x})(x^{(i)}-\bar{x})^T$, we transform the projection to use centered data: $J=\frac{1}{n}\sum_{i=1}^n((x^{(i)}-\bar{x})\cdot u^{[D]})^2$ ○ Expanding the squared projection: $J=\frac{1}{n}\sum_{i=1}^n((x^{(i)}-\bar{x})\cdot u^{[D]})^\top((x^{(i)}-\bar{x})\cdot u^{[D]})=u^{[D]\top}(x^{(i)}-\bar{x})(x^{(i)}-\bar{x})^\top u^{[D]}=u^{[D]\top}Su^{[D]}$ ○ Here, $u^{[D]}$ is the eigenvector of S with the smallest eigenvalue • $x^{(i)}$ is a column of $X^\top$ • If $A=X^\top$ and SVD of A is $USV^\top$, then B is given by $U^{(j\le d)}$, i.e. the first d columns of U • Then, reconstruction error is given by: $J=\frac{1}{n}\sum_{i=1}^n\|x^{(i)}\|^2-\|B^\top x^{(i)}\|^2=\sum_{i=1}^n\|x^{(i)}\|^2-\frac{1}{d}\sum_{i=1}^d\sigma_d^2$ given SVD Projection Energy

Neural Networks

Form. *Form.* — • Model architecture: ○ Input features (X) ¿ weights (B) ¿ weighted sums (S) ¿ activation functions (φ) ¿ hidden states (H) ¿ weights ¿ ... ¿ hidden states ¿ activation funct. ¿ outputs ○ E.g. ■ Outputs: Probability $p(y|x)$ for each class y ■ Activation: Softmax ensures all P add to 1: $p(y|x)=\frac{\exp(h_y^{(K)})}{\sum_{y'}\exp(h_{y'}^{(K)})}$ ■ Hidden layer: $h^{(K)}=\sigma(w^{(K)}h^{(K-1)})$ ... $h^{(1)}=\sigma(w^{(1)}e(x))$ ■ Concatenated vector of word embeddings: $e(x)=\frac{1}{n}\sum w_i e(w_i)$ ■ Transformation: word embedding $e(w_i)$ ■ Inputs: n words • Neuron (j) in layer $[k]$ given training instance $x^{(i)}[0]$ resp. instance from previous layer $h^{(i)}[k-1]$ given by: $h^{(j)}[1]=\varphi(x^{(i)}[0]\cdot\beta^{(j)}[1])$ resp. $h^{(j)}[k]=\varphi(h^{(i)}[k-1]\cdot\beta^{(j)}[k])$ • Outputs for neurons $1,\dots,j$ in fixed layer (notation for layer omitted below) given by: $H=\varphi(XB)=\varphi(S)$ where • $X\in\mathbb{R}^{n\times m+1}$ (incl. bias term) ○ $B\in\mathbb{R}^{m+1\times j}$ with a weight vector for each neuron in each column (incl. bias term) ○ $S\in\mathbb{R}^{n\times j}$ with the weighted sum (prior to activation) for instance i in neuron j is on the i^{th} row and j^{th} column ○ The activation funct. differs by neuron

Activation functions — • Introduce non-linearities • Can differ by neuron • Sigmoid: ○ $[0,1]$ ○ $\varphi(z)=\sigma(z)=\frac{1}{1+e^{-z}}=\frac{e^z}{e^z+1}$ ○ $\varphi'(z)=\frac{e^{-z}}{(1+e^{-z})^2}$ with max. at 0.25 • Hyperbolic Tangent: ○ $[-1,1]$ ○ $\varphi(z)=tanh(z)=\frac{e^z-e^{-z}}{e^z+e^{-z}}=\frac{1-e^{-2z}}{1+e^{-2z}}$ ○ $\varphi'(z)=1-tanh(z)^2$ • ReLu: ○ $\varphi(z)=max(0,z)$ ○ $\varphi'(z)=1$ if z>0;0 otherwise

Opt. *Opt.* — • Perform *forward pass* with randomly initialized parameters, to calculate loss • Perform *backpropagation*, to calculate gradient: ○ $\frac{\partial L}{\partial\theta}=[\frac{\partial L}{\partial B^{[0]}},\dots,\frac{\partial L}{\partial B^{[output]}}]$ ○ $\frac{\partial L}{\partial B^{[k]}}=\frac{\partial L}{\partial H^{[l]}}\frac{\partial H^{[l]}}{\partial B^{[k]}}=C$ ■ When $l>k+1$, i.e. when going several layers back: $\frac{\partial L}{\partial B^{[k]}}=\frac{\partial L}{\partial H^{[l]}}\frac{\partial H^{[l]}}{\partial S^{[l-1]}}\frac{\partial S^{[l-1]}}{\partial H^{[l-1]}}\frac{\partial H^{[l-1]}}{\partial B^{[k]}}$ ■ When $l=k+1$, i.e. when going one layer back: $\frac{\partial L}{\partial B^{[k]}}=\frac{\partial L}{\partial H^{[k+1]}}\frac{\partial H^{[k+1]}}{\partial S^{[k]}}\frac{\partial S^{[k]}}{\partial B^{[k]}}$ • Perform gradient descent to find best weights

Challenges — • Unstable gradients: ○ Exploding gradients: If gradients are >1, gradients grow bigger and bigger during backpropagation, algorithm diverges ○ Vanishing gradients: If gradients are <1 (resp. parameter θ is <1), gradients approach 0 during backpropagation, algorithm fails to converge. Proof: ■ $h_m=\sigma(\theta h_{m-1}+x_m)$ ■ $\frac{\partial h_{m+k}}{\partial h_m}=\prod_{i=0}^{k-1}\frac{\partial h_{m+k-i}}{\partial h_{m+k-i-1}}=\prod_{i=0}^{k}\theta\times\sigma'(\theta h_{m+k-i-1}+x_{m+k-i})$ ■ $\le\prod_i^k\theta\times0.25=\theta^k\times0.25^k$ since derivative of sigmoid has 0.25 as max. value ■ $\to0$ as $k\to\infty$, since $\theta<1$ • Dying ReLUs: ○ Caused when weights are tweaked such that a neuron becomes negative, causing ReLU activation to output 0 ○ Can happen if β_0 in $x^Tw+\beta_0$ is large and negative ○ Dead neuron cannot be brought back: ■ Let $z=x^Tw+b^{[l]}$. ■ ReLU$(b^{[l]})=0$ if $b^{[l]}\le0$ ■ Then $\frac{\partial l}{\partial z}=0$, $\frac{\partial z}{\partial h}=0$, $\frac{\partial h}{\partial b^{[l]}}=0$ ■ $h_{b^{[l]}}$ is guaranteed to be zero for all inputs if h is dead ■ Then, parameters cannot change and h will remain dead

Requirements for backpropagation — • Weights need to be initialized to different values: If they are initialized to the same constant, each neuron produces the same output (since the constant weight is applied to the input), and then during backpropagation, all neurons receive the same gradient updates ○ Weights initialized to 0: ■ $h=\varphi(xB^{[1]})=0$ where h corresponds to z, φ is the primitive, and x corresponds to the input nodes ■ $y=hB^{[2]}=0$ where y corresponds to z' ■ $\frac{\partial L}{\partial B^{[2]}}=\frac{\partial L}{\partial y}\frac{\partial y}{\partial B^{[2]}}=\frac{\partial L}{\partial y}h=0$, i.e. $B^{[2]}$ is not updated ■ $\frac{\partial L}{\partial B^{[1]}}=\frac{\partial L}{\partial y}\frac{\partial y}{\partial h}\frac{\partial h}{\partial B^{[1]}}=\frac{\partial L}{\partial y}B^{[2]}\frac{\partial h}{\partial B^{[1]}}=0$, i.e. $B^{[1]}$ is not updated ■ Thus, weights will always remain 0 and network will not learn ○ Weights initialized to same constant: ■ $B^{[1]}=B^{[2]}$ ■ $h_1=h_2$ ■ $\frac{\partial L}{\partial B^{[1]}}=\frac{\partial L}{\partial B^{[2]}}$ ■ Thus, weights will always receive same updates, will remain equal, and network will not learn • At least one activation must be non-linear so that the ANN does not collapse into a linear regr. Proof: ○ $y=(((XB_1)B_2\dots)B_N$ ○ Since matrix multiplication is associative, we have $B'=B_1\dots B_N$ and $y=XB'$

Convolutional Neural Networks (CNNs)

Form. *Form.* — • Model architecture: ○ Input: Composed of channels (e.g. R with 3 channels) ○ Convolutional layer: Composed of feature maps ○ Channel: Sublayer in input and output, composed of pixels ○ Feature map: Sublayer in convolutional layer, composed of neurons, each neuron is generated by applying filter to all receptive fields across all sublayers in lower layer, weights and biases shared across all neurons in feature map ○ Receptive field: Group of neurons in lower layer, that single neuron in higher layer is connected to, size $f_h\times f_w$ ○ Filter resp. convolutional kernel: Weights applied to all receptive fields across all sublayers in lower layer, size $K\times K$ ○ Zero padding: Padding applied to retain same dimensions in each layer, size $\frac{f_h-1}{2}$ resp. $\frac{f_w-1}{2}$ ○ Stride: By how many neurons receptive field shifts, size $s_h\times s_w$, if stride >1, spatial dimensions in subsequent layer decrease (convolution), if stride <1, spatial dimensions increase (deconvolution) • Output of neuron in layer n, given previous layer $n-1$: $z_{i,j,k}=b_k+\sum_{fn}\sum_{fw}\sum_{f_n'}x_{i',j',k'}\cdot w_{u,v,k',k}$, i.e. sum of element-wise matrix product over all receptive fields and all feature maps, where ○ $z_{i,j,k}$ is the output of neuron in row i and column j on feature map k in layer n ○ f_n and f_w are dimensions of the receptive field in layer $n-1$ ○ f_n' is the number of feature maps in layer $n-1$ ○ $x_{i',j',k'}$ is the output of neuron in row i' and column j' on feature map k' in layer $n-1$ ○ $i'=i\times\text{stride}_h+u-\text{padding}_h$ and $j'=j\times\text{stride}_w+v-\text{padding}_w$ ○ $w_{u,v,k',k}$ is the connection weight between any neuron on feature map k in layer n and its input at u,v on feature map k' ○ $u,v\in\Delta K$ are possible shifts allowed by kernel • Output of neurons in layer n, given previous layer $n-1$: $z_k=b_k+\sum_{fn}\sum_{fw}\sum_{f_n'}W_{k',k}X_{k'}$ • Output size in layer n, given previous layer $n-1$: $H'=\frac{H+2p-K}{\text{stride}_h}+1$ and $W'=\frac{W+2p-K}{\text{stride}_w}+1$

Recurrent Neural Networks (RNN)

Form. Output of neuron in layer n: $Y_t=\phi(X_tW_x+H_{t-1}W_y+b)V=\phi(\begin{bmatrix}X_t\\H_{t-1}\end{bmatrix}W+b)V$

Opt. *Backpropagation through time*: • $\nabla_{W_h}L\propto\sum_{k=1}^t(\prod_{i=k+1}^t\frac{\partial h_i}{\partial h_{i-1}})\frac{\partial h_k}{\partial W_k}$ • $\frac{\partial h_{i+k}}{\partial h_i}=\prod_{j=0}^{k-1}\frac{\partial h_{i+k-j}}{\partial h_{i+k-j-1}}$

Long-Short-Term Memory (LSTM)

Forget gate: • Serves to decide which information to keep from previous cell state, 1: retain, 0: forget • $f_t=\sigma(w_f\cdot[h_{t-1},x_t]+b_f$ where $w_f=\begin{bmatrix}w_{xf}\\w_{hf}\end{bmatrix}$ $\begin{matrix}\text{= connection weight for } x_t\\\text{= connection weight for } h_{t-1}\end{matrix}$ *Input gate*: • Serves to decide which information will be stored in the cell state • $i_t=\sigma(w_i\cdot[h_{t-1},x_t]+b_i$ where $w_i=\begin{bmatrix}w_{xi}\\w_{hi}\end{bmatrix}$ $\begin{matrix}\text{= connection weight for } x_t\\\text{= connection weight for } h_{t-1}\end{matrix}$ • $\tilde{c}_t=tanh(w_{\tilde{c}}\cdot[h_{t-1},x_t]+b_c$ where $w_{\tilde{c}}=\begin{bmatrix}w_{x\tilde{c}}\\w_{h\tilde{c}}\end{bmatrix}$ $\begin{matrix}\text{= connection weight for } x_t\\\text{= connection weight for } h_{t-1}\end{matrix}$ *Cell state*: • Serves to update old cell state into new cell state • $c_t=f_t\otimes c_{t-1}+i_t\otimes\tilde{c}_t$ where $\otimes$ is element-wise multiplication *Output gate*: • Serves to output h_t for next time step, which is a filtered version of cell state c_t • $o_t=\sigma(w_O\cdot[h_{t-1},x_t]+b_o$ where $w_O=\begin{bmatrix}w_{xo}\\w_{ho}\end{bmatrix}$ $\begin{matrix}\text{= connection weight for } x_t\\\text{= connection weight for } h_{t-1}\end{matrix}$ • $h_t=o_t\otimes tanh(c_t)$

Attention

Approach • Steps: ○ Generate three sets of re-weighted embeddings: ■ $Q=EW^q$ resp. $q_i=e_iW^q$ □ E $(m\times h)$ □ W_q $(h\times d_k)$ □ Q $(m\times d_k)$ □ q_i is row vector $(1\times d_k)$ □ e_i is row vector $(1\times h)$ ■ $K=EW^k$ resp. $k_i=e_iW^k$ □ E $(n\times h)$ □ W_k $(h\times d_k)$ □ K $(n\times d_k)$ □ k_i is row vector $(1\times d_k)$ □ e_i is row vector $(1\times h)$ ■ $V=EW^v$ resp. $v_i=e_iW^v$ where □ E $(n\times h)$ □ W_v $(h\times d_v)$ □ V $(n\times d_v)$ □ v_i is row vector $(1\times d_v)$ □ e_i is row vector $(1\times h)$ ○ Compute similarity matrix: $A=\sigma(\frac{QK^\top}{\sqrt{d_k}})$ in $(m\times n)$ resp. $\alpha_t=\sigma(\frac{q_tK^\top}{\sqrt{d_k}})$ resp. $\alpha_{ti}=\frac{exp(q_t\cdot k_i)}{\sum_{i'}exp(q_t\cdot k_{i'})}$ ■ $\sum_i\alpha_{ti}=1$ ■ $\alpha_{ti}\ge0$ ○ Compute attention-weighted embedding matrix: $Z=AV$ in $(m\times d_v)$ resp. $z_t=\alpha_tV=\sum_i\alpha_{ti}v_i$ • In cross-attention: ○ Attention for decoder-encoder alignment ○ Q is generated with decoder input during training with m ○ V,K are generated with encoder outputs during training with n • In self-attention: ○ Attention for encoder resp. decoder inputs ○ Q,V,K are generated with input during training with all either n or m ○ In masked self-attention: ■ We first calculate $P=QK^\top$ where masked elements (e.g. states with time $\ge m$ in decoder) are set to $-\infty$ ■ $S=\sigma(\frac{P}{\sqrt{d_k}})$ • In multi-head attention: ○ Creates multiple sets of Q,K,V and calculates attention correspondingly ○ Concatenates generated matrices Z ○ Multi Head Attention$Z$= Concat$(Z_{head_1},\dots,Z_{head_h})W_O+b_O$ where ■ Concat$(\dots)$ in $(m\times(n\times n_{heads}))$ ■ W_O in $((n_{heads}\times n)\times d_v)$ ■ b_O in $1\times d_v$)

Cookbook

Orthogonality • Find orthonormal basis for subspace spanned by vectors v_i: ○ Perform Gram Schmidt orthogonalization on v_i • Find orthogonal complement for subspace spanned by vectors v_i: ○ Orthogonal complement is a vector u of same dimensionality as basis of subspace ○ Set up equation system $v_i\cdot u=0$ and solve for u • Find projection matrix to subspace spanned by vectors v_i: ○ Find orthonormal basis b_i ○ Calculate matrix $C=BB^\top$ • Find projection matrix to orthogonal complement for subspace spanned by vectors: ○ Calculate $I-C$ ○ Compute projection x_P of vector x to subspace spanned by vectors v_i: ○ Via orthogonality principle: Set up equation system of $(x-x_P)\cdot v_i=(x-(\alpha_i\times v_i))\cdot v_i=0$, solve for α, and calculate $x_P=\sum_i(\alpha_i\times v_i)$ ○ Via orthonormal projection theorem: If b_i is orthonormalized version of v_i, calculate $x_P=\sum_i(x\cdot b_i)b_i$

LSE via Orthogonality • To do: Find parameters h to minimize $||Ah-x||$. Aim: Estimate $\hat{x}$ as Ah. Observe: A and x. ○ Set up equation system $A^\top Ah=A^\top x$ and solve for h • To do: Find minimum norm solution to above problem. ○ Via column space: ■ Express h as vector in column space of A^T: $h=A^T\alpha$ where α is a scaling vector ■ With this, set up equation system $Ah=AA^T\alpha=x$ and solve for α ■ Plug α back into $h=A^T\alpha$ and calculate h ○ Via SVD: ■ Note that h is $A^\# x$ ■ Depending on if A has linearly independent rows or columns, calculate $A^\#$ ■ Plug back $A^\#$ into $A^\# x$ and calculate h

LMMSE Find parameters: • To do: Find parameters a and b to minimize $\mathbb{E}[(\hat{X}-X)^2]$ or to enforce orthogonality $\mathbb{E}[(\hat{X}-X)Y]=0$. Aim: Estimate $\hat{X}$ as aY_1+bY_2 with $Y_2=1$ and $Y_1=Y$. Observe: $Y=cX+Z$. • Set up matrix notation $\mathbb{E}[Y_1^2]h=\mathbb{E}[XY_1]$ • If possible, replace Y_1 with $X+Z$ as well as the other variables with whatever information is given. Consider e.g. PDF, independent variables, variance for zero-mean variables, correlation as decomposition of variance and mean, etc. • Set up equation system and solve for a and b Find estimate: • To do: Evaluate $\mathbb{E}[\hat{X}|X=x]$ • $\mathbb{E}[\hat{X}|X=x]=\mathbb{E}[a(cX+Z)+b|X=x]=acX+b$ • Plug in found parameters a and b

Linear regression • To do: Find parameters h_0 and h_1. Aim: Estimate x as h_0+h_1x. Observe: Pairs of points x,y. • Set up equation system $A^\top Ah=A^\top x$ and solve for h, i.e. OLSE

Col and row rank • For 2×2 or diagonal matrix: Check determinant • For other matrices: ○ Check if one column / row is scalar multiple of another column / row: If yes, linearly dependent ○ For non-zero vectors: Check if inner product column / row equals 0: If yes, linearly independent ○ Matrices with all 0 columns / rows are always linearly dependent

SVD • Find SVD of a matrix A $(n\times m)$ with orthogonal rows: ○ Derive orthogonal square matrix V from A: ■ Orthonormalize rows in A by dividing each row by its norm ■ Verify columns are orthonormal ■ Make matrix square $(m\times m)$ ■ If column is all zero, set it to 1 on diagonal, elsewhere to 0 ■ $=V^\top=(m\times m)$ ○ Derive scaling matrix S from A: ■ Has norm of each row in A on diagonal ■ If extra columns are added, pad with 0 ■ $=S=(n\times m)$ ○ Prepend with identity matrix: ■ $=U=(n\times n)$ ○ $=USV^\top$

MAP and MLE • Determine MAP estimate from empirical data table: ○ Find posterior $P(X|Y)$ and, for each Y, select X yielding highest posterior value • Determine MLE estimate from empirical data table ○ Find likelihood $P(Y|X)$ and, for each Y, select X, under which that Y is likelier • Determine confidence set for a given probability threshold θ: ○ Given set $J_\theta(Y=y)=\{X=x\}$, the term $P(X\in J_\theta(Y)|X=x)\ge\theta$ means: When we observe $Y=y$, we are at least θ% confident that $X=x$ ○ Find likelihood $P(Y|X)$ ○ Go through each discrete X ○ Construct a set J of values of y, such that $\sum_y P(y|X=x)\ge\theta$ ○ For each X, return the smallest possible set of y values (starting with y value contributing the most) such that $P(y\in J|X=x)\ge\theta$ ○ For each Y, write $J_\theta(Y)=\{x \text{ values}\}$

Equation systems • Set free variables (resulting in true = true) to a beliebige variable t • Express other variables in terms of t