For EPFL CS433 ML, ©Arnout Devos

1 Regression

1.1 Linear Regression

$f(\mathbf{x_n}) := w_0 + \sum_{j=1}^{D} w_j x_{nj} = \tilde{\mathbf{x}}_n^T \mathbf{w}$

If $D > N$ the task is under-determined (more dimensions than data) → regularization.

2 Cost functions

MSE = $\frac{1}{N}\sum_{n=1}^{N}[y_n - f(\mathbf{x_n})]^2$ outliers :(

MAE = $\frac{1}{N}\sum_{n=1}^{N}|y_n - f(\mathbf{x_n})|$ outliers :|

2.1 Convexity

$f(\lambda \mathbf{u} + (1-\lambda)\mathbf{v}) \le \lambda f(\mathbf{u}) + (1-\lambda) f(\mathbf{v})$ with $\lambda \in [0;1]$ and $u, v \in$convex set$\mathcal{C}$.

Strictly convex function: unique global minimum w^*. Function always above its linearization: $\mathcal{L}(u) \ge \mathcal{L}(w) + \nabla\mathcal{L}(w)^T(u-w) \forall u, w$.

Set is convex iff line between any two points of $\mathcal{C}$ lies in $\mathcal{C}$: $\theta u + (1-\theta)v \in \mathcal{C}$

3 Optimization

Gradient $\nabla\mathcal{L} := \left[\frac{\partial\mathcal{L}(\mathbf{w})}{\partial w_1} \quad \dots \quad \frac{\partial\mathcal{L}(\mathbf{w})}{\partial w_D}\right]$

3.1 Grid Search $\mathcal{O}(\prod_{i=1}^{D}|W|_i \times N)$

no guarantee (local) optimum close

3.2 Gradient descent $\mathcal{O}(N \times D)$

$\mathbf{w^{(t+1)}} = \mathbf{w^{(t)}} - \gamma\nabla\mathcal{L}(\mathbf{w^{(t)}})$. ill-cond :(

GD - Linear Regression MSE

$\mathcal{L}(\mathbf{w}) = \frac{1}{2N}(\mathbf{y} - X\mathbf{w})^T(\mathbf{y} - X\mathbf{w}) \to$

$\nabla\mathcal{L}(\mathbf{w}) = -\frac{1}{N}X^T(\mathbf{y} - X\mathbf{w})$. Cost: $O_{err} = 2ND + N$ and $O_w = 2ND + D$.

3.3 SGD $\mathcal{O}(D)$

$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_n(\mathbf{w})$ with update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \gamma\nabla\mathcal{L}_n(\mathbf{w}^{(t)})$.

3.4 Mini-batch SGD $\mathcal{O}(|B| \times D)$

$\mathbf{g} = \frac{1}{|B|}\sum_{n\in B}\nabla\mathcal{L}_n(\mathbf{w}^{(t)})$ with update $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \gamma\mathbf{g}$.

3.5 Subgradient at w

$\mathbf{g} \in \mathbb{R}^D$ with $\mathcal{L}(u) \ge \mathcal{L}(w) + \mathbf{g}^T(u-w)$.

3.6 Projected SGD $\mathcal{P}_\mathcal{C}(w') = \arg\min_{v\in\mathcal{C}}\|v - w'\|$

$\mathbf{w}^{(t+1)} = \mathcal{P}_\mathcal{C}[\mathbf{w}^{(t)} - \gamma\nabla\mathcal{L}(\mathbf{w}^{(t)})]$

3.7 Newton's method $\mathcal{O}(ND^2 + D^3)$

2nd order, !cheap, faster convergence

$w^{(t+1)} = w^{(t)} - \gamma^{(t)}(H^{(t)})^{-1}\nabla\mathcal{L}(w^{(t)})$

3.8 Optimality conditions

Necessary: $\nabla\mathcal{L}(\mathbf{w}^*) = 0$, Sufficient: Hessian PSD $\mathbf{H}(\mathbf{w}^*) := \frac{\partial^2\mathcal{L}(\mathbf{w}^*)}{\partial w \partial w^T}$

4 Least Squares

4.1 Normal Equation

$X^T(\mathbf{y} - X\mathbf{w}) = 0 \Rightarrow$ $\mathbf{w}^* = (X^TX)^{-1}X^T\mathbf{y}$ and $\hat{\mathbf{y}}_\mathbf{m} = \mathbf{x_m^T}\mathbf{w}^*$

Gram matrix $\in \mathbb{R}^{D\times D}$ invertible iff $rank(X) = D$ (else use $X = USV^T \in \mathbb{R}^{N\times D}$ to get pseudo-inverse $\mathbf{w}^* = V\tilde{S}U^Ty$ with $\tilde{S}$ pseudo-inverse of S: $\tilde{\sigma_i} = 1/\sigma_i, \forall\sigma_i \ne 0$). cost$(A^{-1}) = \mathcal{O}(N^3)$

5 Likelihood

Probabilistic model $y_n = \mathbf{x_n}^T\mathbf{w} + \epsilon_n$. Probability of observing the data given a set of parameters + inputs: $p(\mathbf{y}|X, \mathbf{w}) = \prod_n p(y_n|\mathbf{x_n}, \mathbf{w}) = \prod_n \mathcal{N}(y_n|\mathbf{x_n}^T\mathbf{w}, \sigma^2)$

Maximizing log-likelihood (=min_w MSE) $\mathcal{L}_{LL} = -\frac{1}{2\sigma^2}\sum(y_n - x_n^Tw)^2 + cst$.

6 Regularization

6.1 Ridge Regression $\mathcal{O}(D^3 + ND^2)$

$\mathcal{L}(\mathbf{w}) = \frac{1}{2N}(\mathbf{y} - X\mathbf{w})^T(\mathbf{y} - X\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2 \quad\to\quad \mathbf{w}^*_{\mathbf{ridge}} = (X^TX + 2N\lambda I)^{-1}X^T\mathbf{y}$

Can be considered a MAP estimator : $\mathbf{w}^*_{\mathbf{ridge}} = argmin_w - log(p(w|X, y))$

6.2 Lasso regularizer

Sparse solution. $\mathcal{L}(w) = \frac{1}{2N}(y - Xw)^T(y - Xw) + \lambda\|w\|_1$

7 Model Selection, eg crossval → λ

$P\left[|L_\mathcal{D} - L_{S_{test}}|) \ge \sqrt{\frac{(b-a)^2 ln(2/\delta)}{2|S_{test}|}}\right] \le \delta$

More data points ($|S_{test}| \uparrow$) = more confident close to true loss. $a \le L \le b$

7.1 Bias-Variance = vary train data

Small dim: large bias, small var. Large dim: small bias, large var. Error for the val set compared to the emp distr of the data $\propto \sqrt{ln(|\Omega|)}/\sqrt{|V|}$.

8 Classification

8.1 Optimal

$\hat{y}(\mathbf{x}) = argmax_{y\in\mathcal{Y}}p(y|\mathbf{x})$

8.2 Logistic regression

$\sigma(z) = \frac{e^z}{1+e^z}$ to limit the predicted values $y \in [0;1]$ ($p(1|\mathbf{x}) = \sigma(\mathbf{x}^T\mathbf{w})$ and $p(0|\mathbf{x}) = 1 - \sigma(\mathbf{x}^T\mathbf{w})$). Decision wrt 0.5.

Likelihood

$p(y|X, w) = \prod p(y_n|x_n)$

$= \prod_{n:y_n=0} p(0|x_n)\dots\prod_{n:y_n=K} p(K|x_n)$

$= \prod_k\prod_n[p(y_n = k|x_n, w)]^{\tilde{y}_{nk}}$

where $\tilde{y}_{nk} = 1$ if $y_n = k$.

For binary classification

$p(y|X, w) = \prod_{n:y_n=0} p(0|x_n)\dots\prod_{n:y_n=1} p(1|x_n)$

$= \prod_n \sigma(x_n^Tw)^{y_n}[1 - \sigma(x_n^Tw)]^{1-y_n}$

Loss

$\mathcal{L}(w) = \sum_{n=1}^{N} ln(1 + exp(x_n^Tw)) - y_nx_n^Tw$ which is convex in w.

Gradient

$\nabla\mathcal{L}(w) = \sum_{n=1}^{N} x_n(\sigma(x_n^Tw) - y_n) = X^T[\sigma(Xw) - y]$ (no closed-form).

Hessian $H(w) = X^TSX$, $S_{nn} = \sigma(x_n^Tw)[1 - \sigma(x_n^Tw)]$ (cvx/Newton 3.7).

If data lin sep: $w* \to \infty$ so regularize

8.3 Exponential family

General form

$p(y|\eta) = h(y)exp[\eta^T\phi(y) - A(\eta)]$

Cumulant

$A(\eta) = ln[\int_y h(y)exp[\eta^T\phi(y)]dy]$

$\nabla A(\eta) = \mathbb{E}[\phi(y)] = \mu = g^{-1}(\eta)$

$\nabla^2 A(\eta) = \mathbb{E}[\phi\phi^T] - \mathbb{E}[\phi]\mathbb{E}[\phi^T]$

Link function

$\eta = g(\mu) \Leftrightarrow \mu = g^{-1}(\eta)$

$\eta_{gaussian} = (\mu/\sigma^2, -1/2\sigma^2)$; $\eta_{poisson} = ln(\mu)$; $\eta_{bernoulli} = ln(\mu/1-\mu)$

$\eta_{general} = g^{-1}(\frac{1}{N}\sum_{n=1}^{N}\phi(y_n))$

GLM: scalar $\phi(y)$, $\eta_n = \boldsymbol{x}_n^T\boldsymbol{w}$, see 8.2

$\nabla_w\mathcal{L}(w) = X^T[g^{-1}(Xw) - \phi(y)] = 0$

8.4 Nearest Neighbor, best low dim

8.4.1 k-NN

$f_{S_{tr},k}(x) = \frac{1}{k}\sum_{n:x_n\in nbh_{S_{tr},k(x)}} y_n$. Pick odd k so clear winner. Large $k \to$ large bias + small variance (inv.)

Error bound, opt Bayes f^*

$\mathbb{E}[\mathcal{L}_{St}] \le 2\mathcal{L}_{f^*} + 4c\sqrt{d}N^{-1/d+1}$

Curse: cst Loss: $N = (1/\alpha)^{d+1}, \alpha << 1$

8.5 Support Vector Machines (SVM)

Logistic regression with hinge loss : $min_w\sum_{n=1}^{N}[1 - y_nx_n^Tw]_+ + \frac{\lambda}{2}\|w\|^2$ where $y \in [-1;1]$ the label and $hinge(x) = max\{0, x\}$. Convex but not differentiable so need subgradient.

Duality: $\mathcal{L}(w) = max_\alpha G(w, \alpha)$. Primal SVM $min_w max_{\alpha\in[0,1]^N}\sum\alpha_n(1 - y_nx_n^Tw) + \lambda/2\|w\|^2$ is diff + cvx. Can switch (=dual) max and min when convex in w and concave in α. Simpler form:

$w(\alpha) = \frac{1}{\lambda}\sum\alpha_ny_nx_n = \frac{1}{\lambda}X^Tdiag(y)\alpha$ which plugging into primal yields:

$max_{\alpha\in[0,1]^N}\alpha^T\mathbf{1} - 1/2\lambda\alpha^TYXX^TY\alpha$

The solution is sparse ($\alpha_n = 0$ correct side, $\alpha_n \in (0, 1)$ on margin, $\alpha_n = 1$ inside margin/wrong side). Coord ascent on α_n

8.6 Kernel Ridge Regression

From duality $w^* := X^T\alpha^*$ where $\alpha^* := (K + \lambda I_N)^{-1}y$ and $K = XX^T = \phi^T(x)\phi(x) = \kappa(x, x')$ (needs to be PSD and symmetric). $\mathcal{O}(N^3 + DN^2)$

9 Unsupervised Learning

9.1 K-means clustering

$min\ \mathcal{L}(z, \mu) = \sum_n^N\sum_k^K z_{nk}\|x_n - \mu_k\|_2^2$ with $z_{nk} \in \{0, 1\}$ (unique assignments: $\sum_k z_{nk} = 1$).

Algorithm (Coordinate Descent z, μ)

1. $\forall n, z_n = \begin{cases} 1 \text{ if } k = argmin_j\|x_n - \mu_j\|^2 \\ 0 \text{ otherwise} \end{cases}$

2. $\forall k$ compute $\mu_k = \sum_n z_{nk}x_n / \sum_n z_{nk}$

Probs:cost,spher+hard clusters

Probabilistic model

$p(X|\mu, z) = \prod_n \mathcal{N}(x_n|\mu_k, I)$

$= \prod_n\prod_k \mathcal{N}(x_n|\mu_k, I)^{z_{nk}}$

9.2 Gaussian Mixture Models

$p(X, z|\mu, \Sigma, \pi) = \prod_n(x_n|z_n, \mu_k, \Sigma_k)p(z_n|\pi) = \prod_n\prod_k[\mathcal{N}(x_n|\mu_k, \Sigma_k)]^{z_{nk}}\prod_k[\pi_k]^{z_{nk}}$

where $\pi_k = p(z_n = k)$

Marginal likelihood: z_n latent variables => factored out of likelihood $p(x_n|\theta) = \sum_k\pi_k\mathcal{N}(x_n|\mu_k, \Sigma_k)$.

nb params with z (D,K«N): $\mathcal{O}(N)$, marg out z: $\mathcal{O}(D^2K)$.

9.3 EM

9.3.1 GMM

Intialize $\mu^{(0)}, \Sigma^{(0)}, \pi^{(0)}$.

1. E-step: Compute the assignments. $q_{kn}^{(t)} := \frac{\pi_k^{(t)}\mathcal{N}(x_n|\mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_k^K\pi_k^{(t)}\mathcal{N}(x_n|\mu_k^{(t)}, \Sigma_k^{(t)})}$ (2. Compute Marginal Likelihood) 3. M-step: Update

$\mu_k^{(t+1)} = \frac{\sum_n q_{nk}^{(t)}x_n}{\sum_n q_{nk}^{(t)}}$ $\pi_k^{(t+1)} = \frac{1}{N}\sum_n q_{nk}^{(t)}$

$\Sigma_k^{(t+1)} = \frac{\sum_n q_{nk}^{(t)}(x_n - \mu^{(t+1)})(x_n - \mu^{(t+1)})^T}{\sum_n q_{nk}^{(t)}}$

9.3.2 General

$\theta^{(t+1)} := argmax_\theta\sum_n\mathbb{E}_{p(z_n|x_n, \theta^{(t)})}[log\ p(x_n, z_n|\theta)]$

10 Matrix Factorisations

10.1 Prediction

Find $\mathbf{X} \approx \mathbf{W}\mathbf{Z}^T$ where $\mathbf{W} \in \mathbb{R}^{D\times K}$ and $\mathbf{Z} \in \mathbb{R}^{N\times K}$ with $K << D, N$. Large $K \to$ overfitting. If $K \ge max\{D, N\}$ trivial solution ($W = \mathbf{1}_D$ or $Z = \mathbf{1}_N$).

Quality of reconstruction (not jointly convex nor identifiable):

$\mathcal{L}(\mathbf{W}, \mathbf{Z}) := \frac{1}{2}\sum_{(d,n)\in\Omega}[x_{dn} - (\mathbf{W}\mathbf{Z}^T)_{dn}]^2$

$= \sum_{(d,n)\in\Omega} f_{dn}(w, z)$

Regularizer: $\frac{\lambda_w}{2}\|\mathbf{W}\|_{Frob}^2 + \frac{\lambda_z}{2}\|\mathbf{Z}\|_{Frob}^2$

Opt SGD ($\nabla_{w_{d'k}}$ for fixed user d' and $\nabla_{z_{n'k}}$ for fixed item n', $= [x_{dn} - (\mathbf{W}\mathbf{Z}^T)_{dn}]\{z_{nk}(d' = d)), w_{dk}(n' = n))\}$) .

ALS (assume no missing ratings):

$\mathbf{Z}_*^T = (\mathbf{W}^T\mathbf{W} + \lambda_zI_K)^{-1}\mathbf{W}^T\mathbf{X}$

$\mathbf{W}_*^T = (\mathbf{Z}^T\mathbf{Z} + \lambda_wI_K)^{-1}\mathbf{Z}\mathbf{X}^T$

10.2 Text Representation

MF of co-occurence X: row($\mathbf{W}$) is wordvec, row($\mathbf{Z}$) is *context* wordvec.

GloVe $f_{dn} = min\{1, (\frac{n_{dn}}{n_{max}})^\alpha\}, \alpha \in [0; 1]$ weighted loss factors. Train as MF.

Skipgram/CBOW, f(.) context/word Bin classif to real/fake word pairs.

10.3 FastText superv $f(y_nWZ^Tx_n)$

Doc-Sent, BoW, $x_n \in \mathbb{R}^{|V|}$=sent, f lin

11 Dimensionality reduction

11.1 SVD (no missing entries)

$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, with $\mathbf{X} : D \times N$, ($\mathbf{U} : D \times D$, $\mathbf{V} : N \times N$) orthonormal, $\mathbf{S} : D \times N$ diag PSD, s_i in desc ord ($s_1 \ge \dots \ge s_D \ge 0$).

Reconstruction

$\|\mathbf{X} - \hat{\mathbf{X}}\|_F^2 \ge \|\mathbf{X} - \mathbf{U}_K\mathbf{U}_K^T\mathbf{X}\|_F^2 = \sum_{i\ge K+1} s_i^2$

$\forall$ rank-K matrix $\hat{\mathbf{X}}$ (i.e. *compress data by project onto left sing vectors.*)

Truncated SVD: $\mathbf{U}_K\mathbf{U}_K^T\mathbf{X} = \mathbf{U}\mathbf{S}_K\mathbf{V}^T$

Application to MF: $\mathbf{U} = \mathbf{W}, \mathbf{S}\mathbf{V}^T = \mathbf{Z}^T$.

Rec. limited by the rank-K of W,Z.

11.2 PCA = rank-K SVD

Decorrelate the data. Empirical cov before: $N\mathbf{K} = \mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{S}_D^2\mathbf{U}^T$. After $\tilde{\mathbf{X}} = \mathbf{U}^T\mathbf{X}$: $N\tilde{\mathbf{K}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T = \mathbf{S}_D^2$ (pure diag = components are uncorrelated). Pitfalls: not invariant under scalings.

12 Neural Networks

The output at the node j in layer l is $x_j^{(l)} = \phi\left(\sum_i w_{i,j}^{(l)} x_i^{(l-1)} + b_j^{(l)}\right)$

12.1 Representation power 1 layer

Error $\leq \frac{(2Cr)^2}{n}$ where C smoothness bound, n #nodes. Approx any sufficiently smooth 2D func on bounded domain (*on avg* σ act, *pointw* ReLU).

12.2 Learning

Problem not convex, but SGD stable. Backpropagation: Let $\mathcal{L}_n = (y_n - f^{(L+1)} \circ \cdots \circ f^{(1)}(\mathbf{x}_n^{(0)}))^2$.

Forward pass

$\mathbf{x}^{(0)} = \mathbf{x}_n$. For $l = 1, \ldots, L+1$
$\mathbf{z}^{(l)} = (\mathbf{W}^{(l)})^T \mathbf{x}^{(l-1)} + \mathbf{b}^{(l)}$, $\mathbf{x}^{(l)} = \phi(\mathbf{z}^{(l)})$

Backward pass

$\delta^{(L+1)} = -2(y_n - \mathbf{x}^{(L+1)})\phi'(\mathbf{z}^{(L+1)})$ and $\forall l : \delta^{(l)} = (\mathbf{W}^{(l+1)}\delta^{(l+1)}) \odot \phi'(\mathbf{z}^{(l)})$

Final pass

$\frac{\partial \mathcal{L}_n}{\partial w_{i,j}^{(l)}} = \delta_j^{(l)} \mathbf{x}_i^{(l-1)}$, $\frac{\partial \mathcal{L}_n}{\partial b_j^{(l)}} = \delta_j^{(l)}$, $\delta_j^{(l)} = \frac{\partial \mathcal{L}_n}{\partial z_j^{(l)}}$

12.3 Activations

sigmoid $\phi(x) = 1/(1+e^{-x}) = 1 - \sigma(x)$, tanh $\frac{e^x + e^{-x}}{e^x + e^{-x}} = 2\phi(2x) - 1$, ReLU, Leaky ReLU ($max\{\alpha x, x\}$).

12.4 Convolutional Neural Nets

Filter f: $x^{(1)}[n,m] = \sum_{k,l} f[k,l] x^{(0)}[n-k, m-l]$. Filter local so no fully connected. Same filter at every position: *weight sharing*. Learning: backprop different weights, sum grads shared weights. Per layer k <<< K: params $\mathcal{O}(kK) = \mathcal{O}(K)$ *else* FFT $\mathcal{O}(L\log(L), L \geq N + K - 1$

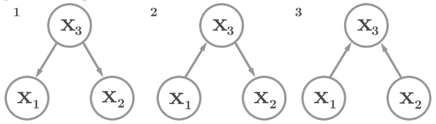
12.5 Overfitting

Adding regularization equivalent to weight decay (by $(1 - \eta\lambda)$). Can also use dataset augmentation, dropout.

13 Graphical Models

13.1 Bayes Nets

$p(X_1, \ldots, X_D) = p(X_1)p(X_2|X_1)\ldots p(X_D|X_1, \ldots, X_{D-1})$. Node is random variable, directed edge from X_j to X_i if X_j appears in the conditioning $p(X_i|\ldots, X_j, \ldots)$. Graph *acyclic* (must). Conditional independence: $p(X,Y) = p(X)p(Y)$ or given Z $p(X,Y|Z) = p(X|Z)p(Y|Z)$.



X3 is tail-to-tail X3 is head-to-tail X3 is head-to-head

1. $p(x_1, x_2, x_3) = p(x_3)p(x_1|x_3)p(x_2|x_3)$: x_1 and x_2 indep. given x_3
2. $p = p(x_1)p(x_3|x_1)p(x_2|x_3)$: id.
3. $p = p(x_1)p(x_2)p(x_3|x_1, x_2)$: x_1 and x_2 **not** indep. given x_3

$X \to Y$ path blocked by Z if it contains a variable such that either 1. variable is in Z and it is head-to-tail or tail-to-tail. 2. node is head-to-head and neither this node nor any of its descendants are in Z.
X and Y are D-sep. by Z iff every path $X \to Y$ is blocked by Z.
X conditionally indep. of Y conditioned on the Z if X and Y are D-sep. by Z. Indep. is symmetric.
Markov blanket (MB) of node X_i is the set of parents, children, and co-parents of the node X_i (other parents of its children). $Y \perp X_i | MB, \forall Y \notin MB$

14 Quick maff

Chain rule $h = f(g(w)) \to \partial h(w) = \partial f(g(w))\nabla g(w)$.
Gaussian
$\mathcal{N}(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(y-\mu)^2}{2\sigma^2})$
Multivariate Gaussian $\mathcal{N}(y|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D det(\Sigma)}} exp(-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu))$
Bayes rule $p(x|y) = p(y|x)p(x)/p(y)$
Logit $\sigma(x) = \partial ln[1+e^x]/\partial x$
Naming Joint distribution $p(x,y) = p(x|y)p(y) = p(y|x)p(x)$ where $p(x|y)$ or $p(y|X,w) \to$ likelihood, $p(y)$ or $p(w) \to$ prior, $p(y|x) \to$ posterior, $p(x) \to$ marginal likelihood, $p(w|y,X) \to$ MAP estimator
Marginal Likelihood
$p(\mathbb{X}|\alpha) = \int_\theta p(\mathbb{X}|\theta)p(\theta|\alpha)\, d\theta$
$p(X = x) = \sum_y p(X = x, Y = y) = \sum_y p(X = x \mid Y = y)p(Y = y)$
Posterior probability $\propto$ Likelihood $\times$ Prior. Max over $\mathcal{N}$ is equiv. to min. MSE:
$\beta^*_{MAP} = argmax_\beta p(y|X,\beta)p(\beta) \Leftrightarrow \beta^* = argmin_\beta \mathcal{L}(\beta)$
Identifiable model
$\theta_1 = \theta_2 \to P_{\theta_1} = P_{\theta_2}$

14.1 Algebra

$(PQ + I_N)^{-1}P = P(QP + I_M)^{-1}$, P^{NM}
$\sum_n (y_n - \beta^T \mathbf{x_n})^2 = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$
$\sum_j \beta^2 = \beta^T\beta$
Unit/ortho: $\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I}$, $\mathbf{U}^T = \mathbf{U}^{-1}$ Rotation matrix (same length vector). Jensen's ineq: $log(\sum a) \geq \sum q log(a/q)$

15 Mock Exam Notes

15.1 Normal equation

Unique if *strictly* convex.
$\frac{1}{\sigma_k^2} X(X^T w_k - y_k) + w_k = 0 \Leftrightarrow$
$w_k^* = (\frac{1}{\sigma_k^2} XX^T + I_D)^{-1} \frac{1}{\sigma_k^2} X y_k$

15.2 MAP solution

$\mathcal{L}(w) = \sum_k \sum_n \frac{1}{2\sigma_k^2}(y_{nk} - x_n^T w_k)^2 + \frac{1}{2}\sum_k \|w_k\|_2^2 \to$ Likelihood $p(y|X,w) = \prod_n \prod_k \mathcal{N}(y_{nk}|w_k^T x_n, \sigma_k^2)$ and prior $p(w) = \prod_k \mathcal{N}(w_k|0, I_D)$

15.3 Deriving marginal distribution

$p(y_n|x_n, r_n = k, \beta) = \mathcal{N}(y_n|\beta_k^T \tilde{x}_n, 1)$
Assume r_n follows a multinomial $p(r_n = k|\pi)$. Derive the marginal $p(y_n|x_n, \beta, \pi)$. $p(y_n|x_n, r_n = k, \beta) = \sum_k^K p(y_n, r_n = k|x_n, \beta, \pi) = \sum_k^K p(y_n|r_n = k, x_n, \beta, \pi) \cdot \pi_k = \sum_k^K \mathcal{N}(y_n|\beta_k^T \tilde{x}_n, \sigma^2) \cdot \pi_k$

15.4 MF

$\hat{r}_{um} = \langle \mathbf{v}_u, \mathbf{w}_m \rangle + b_u + b_m$
$\mathcal{L} = \frac{1}{2}\sum_{u\,m}(\hat{r}_{um} - r_{um}) + \frac{\lambda}{2}\Big[\sum_u (b_u^2 + \|\mathbf{v}_u\|^2) + \sum_m (b_m^2 + \|\mathbf{w}_m\|^2)\Big]$. The optimal value for b_u for a particular user u' : $\sum_{u'\,m}(\hat{r}_{u'm} - r_{u'm}) + \lambda b_{u'} = 0$.
Problem jointly convex? Compute $H(\hat{r}(v,w)) = \begin{bmatrix} 2w^2 & 4vw - 2r \\ 4vw - 2r & 2v^2 \end{bmatrix}$ which is not PSD in general.

16 Multiple Choice Notes

16.1 True statements

- Regularization term $\to$ sometimes min to cvx problem.
- k-NN even data not lin sep.
- $max\{0,x\} = \max_{\alpha \in [0,1]} \alpha x$
  $min\{0,x\} = \min_{\alpha \in [0,1]} \alpha x$
- $g(x) = \min_y f(x,y) \Rightarrow g(x) \leq f(x,y)$
- $\max_x g(x) \leq \max_x f(x,y)$
- $\max_x \min_y f(x,y) \leq \min_y \max_x f(x,y)$
- $\nabla_W(\mathbf{x}^T \mathbf{W} \mathbf{x}) = \nabla_W(\sum_{i,j} W_{i,j} x_i x_j) = \mathbf{x}\mathbf{x}^T$
- $\nabla_x(\mathbf{x}^T \mathbf{W} \mathbf{x}) = (\mathbf{W} + \mathbf{W}^T)\mathbf{x}$
- K-means: opt cluster (centers) init $\to$ one step opt representation points (clusters).
- Logistic loss is typically preferred over L_2 loss in classification.
- For optimising a MF of a $D \times N$ matrix, for large D, N : *per iteration, cost(ALS) > cost(SGD)* and *per iteration, SGD cost* $\neq f(D,N)$.
- The complexity of backprop for a nn with L layers and K nodes/layer is $O(K^2 L)$
- One-dimensional CNN with filter/kernel M non-zero terms. Without bias terms, M parameters per layer.

16.2 Convex functions

- $f(x) = x^\alpha, x \in \mathbb{R}^+, \forall \alpha \geq 1 or \leq 0$
- $f(x) = -x^3, x \in [-1, 0]$
- $f(x) = e^{ax}, \forall x, a \in \mathbb{R}$
- $f(x) = ln(1/x), x \in \mathbb{R}^+$
- $f(x) = g(h(x)), x \in \mathbb{R}, g, h$ convex and increasing over $\mathbb{R}$
- $f(x) = ax + b, x \in \mathbb{R}, \forall a, b \in \mathbb{R}$
- $f(x) = |x|^p, x \in \mathbb{R}, p \geq 1$
- $f(x) = xlog(x), x \in \mathbb{R}^+$
- $ln[\sum_k^K e^{t_k}]$

17 Mock2014

17.1 Weighted LS

$\mathcal{L}(\beta) = 1/2 \sum_n w_n(y_n - \beta^T \tilde{\mathbf{x}}_n)^2$

$\partial \mathcal{L}(\beta) = \sum_n w_n(y_n - \beta^T \tilde{\mathbf{x}}_n)\tilde{\mathbf{x}}_n = -\tilde{X}^T W \mathbf{y} + \tilde{X}^T W \tilde{X} \mathbf{B} = 0.$

$w_n > 0 \to$ W pos def $\to \tilde{X}^T W \tilde{X}$ invertible $\to$ unique sol $\beta^* = (\to \tilde{X}^T W \tilde{X})^{-1} \tilde{X}^T W \mathbf{y}$.
prob model :
$p(\mathbf{y}|X, \beta) = \prod_n \mathcal{N}(y_n|\beta^T \tilde{\mathbf{x}}_n, 1/w_n)$.

17.2 Subgradients

$MAE(\mathbf{w}) = 1/N \sum_n |y_n - f(\mathbf{w}, \mathbf{x_n})|$. Use chain rule with subgradient $h(x) = sgn(x)$.
$\nabla \mathcal{L}(\mathbf{w}) = -1/N \sum_n h(y_n - f(\mathbf{w})) \cdot \nabla f(\mathbf{w}, \mathbf{x_n})$. Then update weights.

17.3 Multiple output reg

x_n has dim D but now y_n has dim K. $\mathcal{L}(\mathbf{W}) = \sum_k \sum_n 1/2\sigma_k^2 (y_{nk} - \mathbf{x}_n^T \mathbf{w})^2 + 1/2\sigma_0^2 \sum_k \|\mathbf{w}_k\|^2$. Derive w.r.t. a $\mathbf{w}_k$ to get optimal weights : $1/\sigma_k^2 X^T(X\mathbf{w}_k - \mathbf{y}_k) + 1/\sigma_0^2 \mathbf{w}_k = 0$. Pb is convex in W. $\mathbf{w}_k^* = (1/\sigma_k^2 X^T X + 1/\sigma_0^2 I_D)^{-1} 1/\sigma_k^2 X^T \mathbf{y}_k$.
Prob model (posterior) same answer as **15.2** but with $1/2\sigma_0^2 I_D$ for the prior

17.4 Kernels

Prove symmetry $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ and PSD $t^T K t = \sum_i \sum_j K_{ij} t_i t_j \geq 0 \forall t$

17.5 Mixture of lin reg

$p(y_n|\mathbf{x}_n, r_n = k, \beta) = \mathcal{N}(y_n|\beta_k^T \tilde{\mathbf{x}}_n, 1)$. We define $\mathbf{r}_{nk}$ like $\mathbf{y}_{nk}$ in **17.2**
Likelihood:
$p(y_n|\mathbf{x}_n, \beta, \mathbf{r}_n) = \prod_k [\mathcal{N}(y_n|\beta_k^T \tilde{\mathbf{x}}_n, \sigma^2]^{r_{nk}}$.
LL:
$p(\mathbf{y}|X, \beta, \mathbf{r}) = \prod_n \prod_k [\mathcal{N}(y_n|\beta_k^T \tilde{\mathbf{x}}_n, \sigma^2]^{r_{nk}}$.
For $p(r_n = k|\pi) = \pi_k$:
$p(y_n|\mathbf{x}_n, \beta, \pi) = \sum_k p(y_n, r_n = k|\mathbf{x}_n, \beta, \pi)$
$= \sum_k p(y_n|r_n = k, \mathbf{x}_n, \beta, \pi) \cdot \pi_k$
$= \sum_k \mathcal{N}(y_n|\beta_k^T \tilde{\mathbf{x}}_n, \sigma^2)\pi_k$.
$-logp(\mathbf{y}|X, \beta, \pi)$
$= -\sum_n log \sum_k \mathcal{N}(y_n|\beta_k^T \tilde{\mathbf{x}}_n, \sigma^2) \cdot \pi_k$.
Model is not convex (sum of gaussians). Not identifiable (by permutation of labels).