Information Retrieval in High Dimensional Data
Formula Collection

Lukas Schüttler

February 27, 2024

1 General

1.1 Data Preparation

Nominal Categories - **No** ordering, **Or**dinal Categories: ordering
Num to Cat: Discretization, Cat to Num: Binarization
Text Preparation: Remove HTML, lower ase, Remove punctuation/numbers/common words, split into words

2 Math Basics

i.i.d: independent and identically distributed

Eigenvectors: $Ax = \lambda x$
A matrix $A \in \mathcal{R}^n$ has eigenvectors if A is square and not singular ($\det(A) \neq 0$).

$$A = A^\top \implies \lambda_i \text{ is real} \qquad \text{rank}(A) = n \implies \lambda_i \neq 0 \quad \forall i$$

$$x^\top Ax > 0 \implies \lambda_i > 0 \qquad x^\top Ax \geq 0 \implies \lambda_i \geq 0 \quad \forall i$$

Positive definite: $x^\top Ax > 0 \quad \forall x \neq 0$
Positive semidefinite: $x^\top Ax \geq 0 \quad \forall x \neq 0$
Jacoobi-Matrix: $\mathbf{J}_f(\mathbf{x}) = \dfrac{df(\mathbf{x})}{d\mathbf{x}} = \left[\dfrac{df_i(\mathbf{x})}{dx_j}\right]_{i=1\ldots m; j=1\ldots n}$
$g(x) \circ f(x) \implies \mathbf{J}_{g\circ f}(\mathbf{x}) = \mathbf{J}_g(f(\mathbf{x})) \cdot \mathbf{J}_f(\mathbf{x})$
$g(\mathbf{x}) = \mathbf{W}x \implies \mathbf{J}_{g\circ f}(\mathbf{x}) = \mathbf{W} \cdot \mathbf{J}_f(\mathbf{x})$
Hessian-Matrix: $\mathbf{H}_f(\mathbf{x}) = \dfrac{d^2 f(\mathbf{x})}{d\mathbf{x}^2} = \left[\dfrac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_k}\right]_{\substack{i=1\ldots m;\\ j,k=1\ldots n}}$
Is symmetric - $\mathbf{H}_f(\mathbf{x}) = \mathbf{J}(\nabla f(\mathbf{x}))$

2.1 Projection

Given a subspace $\mathbb{R}^n \subset \mathbb{R}^p$ with the basis $\mathbf{U}$, the orthogonal projector onto R^n is $\mathbf{P} = \mathbf{U}\mathbf{U}^\top$.

$$\mathbf{P}^2 = \mathbf{P} \qquad \mathbf{P}\mathbf{P}^\top = \mathbf{P}^\top\mathbf{P} = \mathbf{I}$$

2.1.1 Orthogonality Principle ($\min \left[\|\ \underline{\mathbf{y}} - \mathbf{X}\underline{\mathbf{t}}\ \|^2\right]$)

$$\mathbf{y} - \mathbf{X}\mathbf{t} \perp \text{range}[\mathbf{X}] \implies \mathbf{y} - \mathbf{X}\mathbf{t} \in \ker[\mathbf{X}^T]$$

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{t}) = \mathbf{0} \implies \mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\mathbf{t}$$

if $N \geq \text{rank}[\mathbf{X}]$ (All columns of $\mathbf{X}$ are independent, $(\mathbf{X}^T\mathbf{X})^{-1}$ exists)

2.2 Statistics

Normal Distribution: $f_X(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\dfrac{(x-\mu)^2}{2\sigma^2}\right)$
Bernoulli Distribution: $f_X(x) = p^x(1-p)^{1-x} \quad x \in \{0, 1\}$
Maximum likelihood estimation

$$L(x;\theta) = \prod_{i=1}^{N} f_{X_i}(x_i;\theta) \qquad l(x;\theta) = \sum_{i=1}^{N} \log f_{X_i}(x_i;\theta)$$

Curse of dimensionality
Given a random vector $\mathbf{X} \in \mathbb{R}^p$ with the i-th element X_i i.i.d with $Pr(X_i^2 \leq \beta) \leq 1$

$$Pr(||X||_2^2 \geq \beta) \geq 1 - Pr(X_1^2 < \beta)^p$$

In a p-dimensional space, N^p samples are needed to achieve similar results as N samples in a one-dimensional space.

2.3 Convexity

Convex set: $\mathcal{C}$, $\mathbf{x}, \mathbf{y} \in \mathcal{C}, t \in [0, 1] : tx + (1-t)y \in C$
Convex function: $f : C \to \mathbb{R}$, $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$
Concave function: $g : C \to \mathbb{R}$, $g(tx + (1-t)y) \geq tg(x) + (1-t)g(y)$
If the Hessian is positive semidefinite e.g all entries $H_{ij} \geq 0$ (second derivative of a function is positive) the function is convex.

2.3.1 Properties:

Given two convex functions f and g, the following functions are also convex:

$$h = \max(f, g) \qquad h = f + g \qquad h = g \circ f \text{ if } g \text{ is non-decreasing}$$

2.4 Non-Linear Optimization

$$\min_{\mathbf{z}} f(\mathbf{z}) \qquad \text{s.t.} \quad c_i(\mathbf{z}) = 0 \quad i \in \mathcal{E} \qquad c_i(\mathbf{z}) \geq 0 \quad i \in \mathcal{I}$$

Lagrange-function: $L(\mathbf{z}, \lambda) = f(\mathbf{z}) - \sum_i \lambda_i c_i(\mathbf{z})$

2.4.1 Karush-Kuhn-Tucker-Conditions

$$\begin{aligned} &\nabla_{\mathbf{z}} L(\mathbf{z}^*, \lambda^*) = 0 && \lambda_i^* c_i(\mathbf{z}^*) = 0 && \\ &\lambda_i^* \geq 0 && c_i(\mathbf{z}^*) \geq 0 && \text{for } i \in \mathcal{I} \\ & && c_i(\mathbf{z}^*) = 0 && \text{for } i \in \mathcal{E} \end{aligned}$$

2.4.2 Lagrangia Duality

$$g(\lambda) = \inf_{\mathbf{z}} L(\mathbf{z}, \lambda) \qquad \max_{\lambda} g(\lambda) \quad \text{s.t.} \quad \lambda_i \geq 0$$

3 Kernel Trick

$\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel function** if the following two conditions are fulfilled (with an arbitrary function $f \in L_2(\mathcal{X})$):
Positive definite: $\int_{\mathcal{X}\times\mathcal{X}} f(\mathbf{x})\kappa(\mathbf{x}, \mathbf{y})f(\mathbf{y})dxdy \geq 0$
Symmetry: $\kappa(\mathbf{x}, \mathbf{y}) = \kappa(\mathbf{y}, \mathbf{x})$

3.1 Common Kernels ($a, c, d \geq 0$ and $\sigma > 0$)

Linear Kernel: $\kappa(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y} + c$
Polynomial Kernel: $\kappa(\mathbf{x}, \mathbf{y}) = (a\mathbf{x}^T\mathbf{y} + c)^d$
Gaussian Kernel: $\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-\dfrac{||\mathbf{x} - \mathbf{y}||^2}{2\sigma^2}\right)$
Exponential Kernel: $\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-\dfrac{||\mathbf{x} - \mathbf{y}||}{2\sigma^2}\right)$
Radial basis function (RBF) Kernel: Gaussian Kernel
Sigmoid Kernel: $\kappa(\mathbf{x}, \mathbf{y}) = \tanh\left(\gamma\mathbf{x}^T\mathbf{y} - \delta\right)$

3.2 Properties

Given the kernels κ_1 and κ_2, $c > 0$ and an arbitrary function f the following combinations are valid kernels:

$$c\kappa_1(\mathbf{x}) \qquad c + \kappa_1(\mathbf{x}) \qquad f(\mathbf{x})f(\mathbf{y})$$

$$\kappa_1(\mathbf{x})\kappa_2(\mathbf{x}) \qquad \kappa_1(\mathbf{x}) + \kappa_2(\mathbf{x})$$

Mercer's Theorem: Let κ be a kernel, then there exists functions ϕ_i and a $\lambda_i \geq 0$ such that:

$$\kappa(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x})\phi_i(\mathbf{y})$$

4 Unsupervised Learning

4.1 K-Means

$$\mathcal{F} = \{f : \mathbb{R}^p \to \{\mathbf{c}_1, \dots, \mathbf{c}_k\} \subset \mathbb{R}^p\} \quad L(\mathbf{X}, f(\mathbf{X})) = ||\mathbf{X} - \mathbf{f}(\mathbf{X})||^2$$

Where $\mathbb{R}^p$ is the feature space. $\mathbf{c}_i$ is the center of cluster i.
Because f maps to a finite set, the volume of the set distribution is zero.
Algorithms: Lloyd's algorithm, MaxQueen's algorithm

4.2 Principle Component Analysis

$$\mathcal{F} = \left\{f : \mathbb{R}^p \to \mathbb{R}^k \subset \mathbb{R}^p\right\} \quad L(\mathbf{X}, f(\mathbf{X})) = ||\mathbf{X} - \mathbf{U}_k\mathbf{U}_k^\top\mathbf{X}||_2^2$$

Where $\mathbb{R}^p$ is the feature space. And $f(\mathbf{X}) = \mathbf{U}_k^\top\mathbf{X}$ is a (orthogonal) projection on to the subspace $\mathbb{R}^k$.
Principle components: $\mathbf{S} = \mathbf{U}_k^\top\mathbf{X} = \mathbf{\Sigma}_k\mathbf{V}_k^\top$
New data: $\mathbf{s}_i = \mathbf{U}_k^\top\mathbf{y}_i = \mathbf{\Sigma}_k^{-1}\mathbf{V}_k^\top\mathbf{X}^\top\mathbf{y}_i$

4.2.1 Kernel PCA

$$\tilde{\mathbf{K}} = \mathbf{HKH} = \mathbf{V\Sigma}^2\mathbf{V}^\top \qquad \mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \qquad \mathbf{S} = \mathbf{\Sigma V}^\top$$

$$\mathbf{s}_i = \mathbf{\Sigma}_k^{-1}\mathbf{V}_k^\top\mathbf{k} \qquad k = H\left([\kappa(x_1, y_i), \dots, \kappa(x_n, y_i)]^\top - \frac{1}{n}\mathbf{K1}\right)$$

5 Supervised Learning

Expected Prediction Error: $\mathrm{EPE(f)} = \mathrm{E}\left[L(Y, f(X))\right]$

$$f(x) = \mathrm{argmin}_{f\in\mathcal{F}}\,\mathrm{EPE}(f) \approx \mathrm{argmin}_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n L(y_i, f(x_i)) = \hat{f}(x)$$

Quadratic Loss: $L_2(Y, f(X)) = (Y - f(X))^2$
Absolute Loss (l_1-loss): $L_1(Y, f(X)) = |Y - f(X)|$

$$f_2(x) = \mathrm{E}_{Y|X=x}[Y] \qquad f_1(x) = \mathrm{median}_{Y|X=x}[Y]$$

Generalization error: $G_N(f) = \mathrm{EPE}(f) - \frac{1}{N}\sum_{i=1}^N L(y_i, f(x_i))$

5.1 Linear regression

$$\mathcal{F} = \left\{f(\mathbf{x}) = \theta_0 + \sum_{k=1}^p \theta_k x_k \middle| \theta_k \in \mathbb{R}, p \in \mathbb{N}\right\}$$

Where x_k is the kth element of the vector $\mathbf{x}$

5.2 K-nearest Neighbors

$$\hat{f}_k(x) = \frac{1}{k}\sum_{i\in N_k(x)} y_i \quad \forall i \text{ s.t. } x_i \in N_k(x)$$

Where $N_k(x)$ is the set of the k nearest neighbors of x

$$\lim_{N,k\to\infty, \frac{k}{N}\to 0} \hat{f}_k(x) = \mathrm{E}[Y|X = x]$$

5.3 Logistic regression

$$\mathcal{F} = \{f(\mathbf{x}) = \mathbf{w}^\top\mathbf{x} + b \mid \mathbf{w} \in \mathbb{R}^p, b \in \mathbb{R}\}$$

$$L_{0,1}(Y, f(X)) = \begin{cases}1 & \text{if } Yf(X) \leq 0\\ 0 & \text{otherwise}\end{cases} \quad l(Y, f(X)) = \log\left(1 + e^{-Yf(X)}\right)$$

f could be an arbitrary function, but usually f is chosen as defined above.

$$Pr(Y = y|\mathbf{x}) = \exp(-l(y, f(\mathbf{x}))) = \frac{1}{1 + e^{-yf(\mathbf{x})}} = \sigma(yf(\mathbf{x}))$$

5.3.1 Statistical Approach

$$\mathbf{w}_{ML} = \mathrm{argmax}_\mathbf{w}\, L(\mathbf{w}) = \mathrm{argmax}_\mathbf{w}\, Pr(\mathbf{y}|\mathbf{x}, \mathbf{w}) \sim \mathrm{argmin}_\mathbf{w} - \log Pr(\mathbf{y}|\mathbf{x}, \mathbf{w})$$

$$Pr(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^n Pr(y_i|\mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^n \sigma(\mathbf{w}^\top\mathbf{x}_i)^{y_i}(1 - \sigma(\mathbf{w}^\top\mathbf{x}_i))^{1-y_i}$$

$$\mathbf{w}_{t+1} = (\mathbf{XBX}^\top)^{-1}\mathbf{XBr}_t \quad \mathbf{r}_t = \mathbf{X}^\top\mathbf{w}_t - \mathbf{B}^{-1}(\sigma(\mathbf{X}^\top\mathbf{w}_t) - \mathbf{y})$$

$$\mathbf{B} = \mathrm{diag}(\sigma(\mathbf{w}^\top\mathbf{x}_i)(1 - \sigma(\mathbf{w}^\top\mathbf{x}_i))) \qquad H = \nabla_\mathbf{w} L(\mathbf{w}) = \mathbf{XBX}^\top$$

Regularization: $\tilde{l}(y, f(\mathbf{x})) = l(y, f(\mathbf{x})) + \lambda\left(||\mathbf{w}||^2 + b^2\right)$

5.4 Feedforward Neural Network

$$f : \mathbb{R}^p \to \mathbb{R}^r \qquad x \mapsto \sigma_l \circ \varphi_{\mathbf{w}_l} \circ \dots \circ \sigma_1 \circ \varphi_{\mathbf{w}_1}(x)$$

$$\varphi_\mathbf{w} : \mathbb{R}^p \to \mathbb{R}^m \quad \mathbf{x} \mapsto \mathbf{Wx} \qquad \sigma : \mathbf{x} \mapsto [\sigma(x_1), \dots, \sigma(x_m)]^\top$$

Where m is the number of neurons in the layer and p is the number of input features/neurons from the previous layer.
Rectified Linear Unit (ReLU): $\sigma(x) = \max(0, x)$

Update rule: $\mathbf{W}_j \leftarrow \mathbf{W}_j - \alpha\pi_j^{-1}\left(\frac{d}{d\mathbf{W}_j}L(\mathbf{y}_i, f(\mathbf{x}_i))\right)^\top$

Softmax: $\mathbf{x} \mapsto \left(\sum_\mathbf{x}\exp(x_i)\right)^{-1}[\exp(x_1), \dots, \exp(x_C)]^\top$
Cross entropy: $H(p, q) = -\sum_{i=1}^n p_i \log q_i$; $L(f(\mathbf{x}), \mathbf{y}_c) = -log(f(\mathbf{x})_c)$
Where c is the correct class in a multiclass classification problem with C classes

5.5 Support Vector machine

$$\mathcal{H}_{\mathbf{w},b} = \{\mathbf{x} \in \mathbb{R}^p | \mathbf{w}^\top\mathbf{x} - b = 0\} \qquad \delta(\mathbf{x}, \mathcal{H}_{\mathbf{w},b}) = \frac{\mathbf{w}^\top\mathbf{x} - b}{||\mathbf{w}||}$$

$$\mathcal{H}_\pm = \{\mathbf{x} \in \mathbb{R}^p | \mathbf{w}^\top\mathbf{x} - b = \pm1\} \qquad \delta(\mathcal{H}_-, \mathcal{H}_+) = \frac{2}{||\mathbf{w}||}$$

$$\max\frac{2}{||\mathbf{w}||} \text{ or } \min\frac{1}{2}||\mathbf{w}||^2 \quad \text{s.t. } y_i(\mathbf{w}^\top\mathbf{x}_i - b) \geq 1 \quad \forall i \in \{1, \dots, n\}$$

Applying Lagrange duality:

$$\min_{\mathbf{w},b,\lambda\geq0} L(\mathbf{w}, b, \lambda) \quad L(\dots) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^n \lambda_i(y_i(\mathbf{w}^\top\mathbf{x}_i - b) - 1)$$

$$\max_\lambda\left(\sum_i \lambda_i - \frac{1}{2}\lambda^\top\mathbf{H}\lambda\right) \quad \text{s.t.} \quad \lambda_i \geq 0, \sum_i \lambda_i y_i = 0$$

Where $\mathbf{H}$ is defined with $h_{ij} = y_i y_j \mathbf{x}_i^\top\mathbf{x}_j$
If $\lambda_i \neq 0$, then $\mathbf{x}_i$ is a support vector ($\mathbf{x}_i \in \mathcal{H}_+ \cup \mathcal{H}_-$)
KKT-Conditions:

$$\nabla_{(\mathbf{w},b)}L(\mathbf{w}, b, \lambda) = \left[\mathbf{w} - \sum_i \lambda_i y_i \mathbf{x}_i, \quad \sum_i \lambda_i y_i \mathbf{x}_i\right]^\top$$

$$\mathbf{w}^* - \sum_i \lambda_i^* y_i \mathbf{x}_i = 0 \quad \sum_i \lambda_i^* y_i \mathbf{x}_i \quad \lambda_i^*(y_i((w^*)^\top\mathbf{x}_i - b^*)) - 1 = 0$$

5.5.1 Soft margin SVM

$$\min_{\mathbf{w},b,\xi}\frac{1}{2}||\mathbf{w}||^2 + c\sum_{i=1}^n \xi_i \qquad \text{s.t. } y_i(\mathbf{w}^\top\mathbf{x}_i - b) \geq 1 - \xi_i,\ \xi_i \geq 0$$

$$L(\mathbf{w}, b, \xi, \lambda, \mu) = \frac{1}{2}||\mathbf{w}||^2 + c\sum_i \xi_i - \sum_i \lambda_i(y_i(\mathbf{w}^\top\mathbf{x}_i - b) - 1 + \xi_i) - \sum_i \mu_i\xi_i$$

$$\max_\lambda\left(\sum_i \lambda_i - \frac{1}{2}\lambda^\top\mathbf{H}\lambda\right) \qquad \text{s.t.} \quad 0 \leq \lambda_i \leq c, \sum_i \lambda_i y_i = 0$$

$$b^* = \frac{1}{N_{\text{supp}}}\sum_{i\in\text{supp}}\left((\mathbf{w}^*)^\top\mathbf{x}_i - y_i\right)$$

5.5.2 Kernel SVM

All vector products are replaced by the kernel $\kappa(\mathbf{x}, \mathbf{y})$

$$h_{ij} = y_i y_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \qquad \mathbf{w}^\top\mathbf{x}_i \neq \kappa(\mathbf{w}, \mathbf{x}_i)$$

$$b^* = \frac{1}{N_{\text{supp}}}\sum_{i\in\text{supp}}\left(\sum_{j\in\text{supp}}(\lambda_j y_j \kappa(\mathbf{x}_i, \mathbf{x}_j)) - y_i\right)$$