1 Linear Regression

$\mathbf{x} \in \mathbb{R}^d$, $y \in \mathbb{R}$, $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{y} \in \mathbb{R}^n$, $\Phi \in \mathbb{R}^{n\times p}$

$\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n (y_i - \mathbf{w}^T\mathbf{x}_i)^2 = \arg\min_{\mathbf{w}\in\mathbb{R}^d} ||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2$

($\nabla_{\mathbf{w}}||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2 = 2\mathbf{X}^T(\mathbf{X}\widehat{\mathbf{w}}-\mathbf{y}) = \mathbf{0}$)

If $n \geq d$, $\mathbf{X}$ full r.: $\widehat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
If $n < d$, $\mathbf{X}$ full r.: $\widehat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{\dagger}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{y}$

If $n \leq d$ there always exists $\mathbf{w}$ such that $\mathbf{y} = \mathbf{X}\mathbf{w}$

2 Optimization

Closed-form solution of linear regression: $\mathcal{O}(n^3 + nd^2)$

2.1 Gradient Descent

$\mathbf{w}^{t+1} = \mathbf{w}^t - \eta\nabla L(\mathbf{w}^t)$, real step size: $\eta||\nabla L(\mathbf{w}^t)||$

$\lambda_{max} := \lambda_{max}(\mathbf{X}^T\mathbf{X})$ & $\lambda_{min} := \lambda_{min}(\mathbf{X}^T\mathbf{X})$, $\kappa := \frac{\lambda_{max}}{\lambda_{min}}$

$\eta_{opt} = \frac{2}{\lambda_{max}+\lambda_{min}}$ $\quad \eta < 2/\lambda_{max}$ $\quad \rho_{min} = \frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}}$

2.2 Momentum-based Methods & SGD

$\mathbf{w}^{t+1} = \mathbf{w}^t + \alpha(\mathbf{w}^t - \mathbf{w}^{t-1}) - \eta\nabla L(\mathbf{w}^t)$

$|minibatch|$ impacts updates & comp. complexity

2.3 Convexity

0th-order: $f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{x})$
1st-order: $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle\nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle$
2nd-order: $D^2 f(\mathbf{x}) \succeq 0$, or $\mathbf{x}^T D^2 f(\mathbf{x})\mathbf{x} \geq 0$

1. $\alpha f + \beta g$ convex, if f, g convex and $\alpha, \beta \geq 0$
2. $f \circ g$ convex, if f convex and g affine, or f non-decreasing and g convex
3. $max(f,g)$ convex, if f, g convex

Strong Conv.: f least as conv. as quad. f: $D^2 f(\mathbf{x}) \succeq mI$

3 Model evaluation and selection

3.1 Estimation and Generalization Error

Train/Test Error: $\frac{1}{|D_{train/test}|}\sum_{(x,y)\in D_{train/test}} \ell(f(x), y)$

Exp. estimation Error: $\mathbb{E}_X \ell(\widehat{f}_D(X), f^*(X))$
Generalization error: $\mathbb{E}_{X,Y}\ell(\widehat{f}_D(x), Y)$

3.2 K-fold cross-validation (popular K={5,10})

4 Bias-Variance & Ridge/LASSO Regular.

$\mathbb{E}_D[L(\widehat{f}_D; \mathbb{P}_{X,Y})] = \mathbb{E}_{X,Y,D}[(\widehat{f}_D(X) - Y)^2]$

$= \mathbb{E}_{X,D}[(\widehat{f}_D(X) - \mathbb{E}_D[\widehat{f}_D(X)])]^2 \quad \rightarrow Var_D(\widehat{f}_D)$
$+ \mathbb{E}_X[(\mathbb{E}_D[\widehat{f}_D(X)] - f^*(X))^2] \quad \rightarrow Bias_D^2(\widehat{f}_D)$
$+ \mathbb{E}_{X,Y}[(f^*(X) - Y)^2] \quad \rightarrow \sigma^2$

LASSO: $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} ||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_1$
Ridge: $\widehat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} ||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_2^2$

$\widehat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_d)^{-1}\mathbf{X}^T\mathbf{y}$, Optimal λ selected with CV

LASSO induces *sparsity* (many coefficients set to zero)

5 Classification

0-1 loss: $\ell_{0-1}(\text{sign} f_{\mathbf{w}}(\mathbf{x}), y) = \mathbb{I}_{sign f_w(x)\neq y} = \mathbb{I}_{f_w(x)\cdot y<0}$
log. loss: $\ell_{log}(\mathbf{w}) = log(1 + e^{-y f_{\mathbf{w}}(\mathbf{x})}) = log(1 + e^{-y\mathbf{w}^T\mathbf{x}})$

5.1 Max-margin Sol. and Logistic Regression

$\mathbf{w}_{\text{MM}} = \arg\max_{||\mathbf{w}||_2=1} \text{margin}(\mathbf{w}) = \arg\max_{||\mathbf{w}||_2=1} \min_{1\leq i\leq n} y_i\langle\mathbf{w}, \mathbf{x}_i\rangle$

$\mathbf{w}_{\text{SVM}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} ||w||_2$ s.t. $y_i\langle\mathbf{w}, \mathbf{x}_i\rangle \geq 1, \forall i = 1-n$

$\mathbf{w}_{MM} = \mathbf{w}_{SVM}/||\mathbf{w}_{SVM}||_2$

Linear separable data: Logistic regression $\rightarrow \mathbf{w}_{MM}$

Length of projection of $\mathbf{x}$ onto $\mathbf{w}$ $\rightarrow \frac{\mathbf{w}^T\mathbf{x}}{||\mathbf{w}||}$

5.2 Soft-Margin Solution

Hinge loss: $\max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)$

$\min_{\mathbf{w}\in\mathbb{R}^d, \xi\in\mathbb{R}^n} [||w||_2^2 + \lambda\sum_{i=1}^n \max(0, 1 - y_i\mathbf{w}^T\mathbf{x}_i)]$

5.3 Multiclass classification

Train One-vs-rest binary calssifier: $\widehat{y}(\mathbf{x}) = \arg\max_{1\leq k\leq K} \widehat{f}_k(\mathbf{x})$

x-entropy: $\ell_{ce}(\widehat{f}_1(\mathbf{x}), ..., \widehat{f}_K(\mathbf{x}), y) = -log\left(\frac{e^{\widehat{f}_y(\mathbf{x})}}{\sum_{k=1}^K \widehat{f}_k(\mathbf{x})}\right)$

5.4 Evaluation metrics for classifiers

Important class (*null hypothesis*): *negative* labels -1
FP: predict +1, when real -1 (more important then FN)

Precision	$\frac{\#TP}{\#\{\widehat{y}=+1\}}$	FDR (1 - Precision)	$\frac{\#FP}{\#\{\widehat{y}=+1\}}$
Recall (TPR, power)	$\frac{\#TP}{\#\{y=+1\}}$	FPR (type I error)	$\frac{\#FP}{\#\{y=-1\}}$
FNR (type II error)	$\frac{\#FN}{\#\{y=+1\}}$	TNR	$\frac{\#TN}{\#\{y=-1\}}$

F1-*score*: $\frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}$ $\quad$ precision + FDR = 1

Assymetric error function: $c_{\text{FN}} \cdot \text{FNR} + c_{\text{FP}} \cdot \text{FPR}$
When $c_{\text{FN}} < c_{\text{FP}}$, it is more important to control FP

5.5 ROC curve (TPR+FNR=1)(TNR+FPR=1)

τ deacreases $\rightarrow$ bigger TPR & FPR
Ideal AUROC = 1, randomly guessing AUROC = 1/2

6 Kernels

6.1 Kernelization

$\mathbf{w} = \mathbf{\Phi}^T\boldsymbol{\alpha}$, $\boldsymbol{\alpha} \in \mathbb{R}^n$, $\mathbf{\Phi} \in \mathbb{R}^{n\times p}$ $\quad \widehat{\mathbf{w}} = \sum_{i=1}^n = \widehat{\alpha}_i\phi(\mathbf{x}_i)$

$f(\mathbf{x}) = \langle\mathbf{w}, \boldsymbol{\phi}(\mathbf{x})\rangle = \langle\mathbf{\Phi}^T\boldsymbol{\alpha}, \boldsymbol{\phi}(\mathbf{x})\rangle$
$= \sum_{i=1}^n \alpha_i\langle\boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x})\rangle = \sum_{i=1}^n \alpha_i k(\mathbf{x}_i, \mathbf{x})$

$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} ||\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}||$, if $\mathbf{K}$ is invertible, then $\widehat{\boldsymbol{\alpha}} = \mathbf{K}^{-1}\mathbf{y}$

Kernel ridge: $\frac{1}{n}||\mathbf{y} - \mathbf{K}\boldsymbol{\alpha}||_2^2 + \lambda\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}$, $\widehat{\boldsymbol{\alpha}} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$
Eigenvalues: $\lambda_i^2 + \lambda\lambda_i$, $\quad \lambda_i$: EV of $\mathbf{K}$

Memory: For $\boldsymbol{\phi}(\mathbf{x}_i) \in \mathbb{R}^p$ for $i = 1-n \rightarrow \mathcal{O}(np)$
For $k(\mathbf{x}_i, \mathbf{x}_j) \in \mathbb{R}$ for $i, j = 1-n \rightarrow \mathcal{O}(n^2)$

p get big fast: A poly. of deg. m of d-dim. inp., need $p = \binom{m+d}{m} = \mathcal{O}(d^m)$ features, if m»d $h(m) = \mathcal{O}(m^d)$

Num. of computations: from $\mathcal{O}(n^2 d^m)$ to $\mathcal{O}(n^2(d+m))$

6.2 Valid kernel functions

1. k is sym.: $k(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}', \mathbf{x})$ 2. K-Matrix $\mathbf{K}$ is p.s.d.

6.3 Examples of kernels

Inner product of kernel: $k(\mathbf{x}, \mathbf{x}') = g(\langle\mathbf{x}, \mathbf{x}'\rangle)$
$\rightarrow$ polynomial kernel $k(\mathbf{x}, \mathbf{x}') = (1 + \langle\mathbf{x}, \mathbf{x}'\rangle)^m$

RBF kernels: $k(\mathbf{x}, \mathbf{x}') = g(||\mathbf{x} - \mathbf{x}'||)$
$\rightarrow$ α-exponential kernel $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||_p^\alpha}{\tau}\right)$

Gaussian: $\alpha = 2$ & $p = 2$ $\quad$ *Laplacian*: $\alpha = 2$ & $p = 1$

7 Neural Networks

7.1 Activation Function

Sigmoid: $\varphi(z) = \frac{1}{1+\exp(-z)}$ $\quad$ RELU: $\varphi(z) = \max(0, z)$
Hyperbolic tangent: $\varphi = \tanh z = \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)}$

7.2 Universal Approximation Theorem

$f : [0,1]^d \rightarrow \mathbb{R}$, φ the sigmoid, $\rightarrow f$ approx. by finite sum:

$\widehat{f}(\mathbf{x}) = \sum_{j=1}^m w_j^{(2)}\varphi\left(\left(w_j^{(1)}\right)^T\mathbf{x} + w_{j,0}^{(1)}\right)$

7.3 Forward Propagation (inp. lay. $\mathbf{v}^{(0)} = [\mathbf{x}, 1]$)

1. $\mathbf{z}^{(l)} = \mathbf{W}^{(l)}\mathbf{v}^{(l-1)}$ and $\mathbf{v}^{(l)} = [\varphi(\mathbf{z}^{(l)}); 1]$
2. for output layer $f = \mathbf{W}^{(l)}\mathbf{v}^{(l-1)}$

7.4 Back Propagation

for L: - error: $\boldsymbol{\delta}^{(L)} = \nabla_f\ell$
. - gradient: $\nabla_{\mathbf{W}^{(L)}}\ell = \text{diag}\left(\boldsymbol{\delta}^{(L)}\right)\mathbf{1}_K\left(\mathbf{v}^{(L-1)}\right)^T$

for l: - error: $\boldsymbol{\delta}^{(l)} = \text{diag}(\dot{\boldsymbol{\varphi}}(\mathbf{z}^{(l)}))(\mathbf{W}^{(l+1)})^T\boldsymbol{\delta}^{(l+1)}$
. - gradient: $\nabla_{\mathbf{W}^{(l)}}\ell = \text{diag}(\boldsymbol{\delta}^{(l)})\mathbf{1}_{n_l}(\mathbf{v}^{(l-1)})^T$

7.5 Weight initialization

RELU: $\mathcal{N}\left(0, \frac{2}{n_{in}}\right)$, tanh: $\mathcal{N}\left(0, \frac{1}{n_{in}}\right)$ or $\mathcal{N}\left(0, \frac{2}{n_{in}+n_{out}}\right)$

Ensures equal (const.) variance of neurons in each layer.

7.6 Other NN stuff

Dropout: $1-p$ prob. to eliminate unit & freeze $\mathbf{w}$.

To compensate, we multiply $\mathbf{w}$ with p during test time.

7.7 Convolutional Neural Networks

Output image dim: $\left(\frac{n+2p-f}{s}+1\right) \times \left(\frac{n+2p-f}{s}+1\right) \times m$

8 Clustering

8.1 k-Means Clustering and Lloyd's Heuristic

1. $z_i^{(t)} \leftarrow \arg\min_{j=1,...,k} ||\mathbf{x}_i - \boldsymbol{\mu}_j^{(t-1)}||_2, \qquad i = 1, ..., n$
2. $\boldsymbol{\mu}_j^{(t)} \leftarrow \frac{1}{n_j^{(t)}} \sum_{i:z_i^{(t)}=j} \mathbf{x}_i, \qquad j = 1, ..., k$

Converges to *local* opt. Dependend on initialization.

++: Next $\mu_j^{(0)}$ with prob. to dist.^{2} to clos. $\mu_i^{(0)}$ $\mathcal{O}(log(k))$

8.2 Choosing k [regularized loss: $\widehat{R} + \lambda \cdot k$]

Plot cost $\widehat{R}$ against k, identify the *elbow* (or kink) of curve.

9 Principal Component Analysis with k=1

$\mathbf{w}^*, z_1^*, ..., z_n^* = \arg\min_{\mathbf{w}\in\mathbb{R}^d:||\mathbf{w}||_2=1,\ z_1,...,z_n\in\mathbb{R}} \sum_{i=1}^n ||\mathbf{x}_i - z_i\mathbf{w}||_2^2$

Optimal $z_i^* = \mathbf{w}^T\mathbf{x}_i, \qquad \boldsymbol{\Sigma} = \mathbf{X}^T\mathbf{X} = \sum_{i=1}^d \lambda_i \mathbf{v}_i \mathbf{v}_i^T$

$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^d:||\mathbf{w}||_2=1} \sum_{i=1}^n ||\mathbf{x}_i - \mathbf{w}\mathbf{w}^T\mathbf{x}_i||_2^2$

$= \arg\max_{||\mathbf{w}||_2=1} \sum_{i=1}^n \left(\mathbf{w}^T\mathbf{x}_i\right)^2 = \arg\max_{||\mathbf{w}||_2=1} \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$

Sol.: $\mathbf{w}^* = \mathbf{v}_1$, with $(\mathbf{w}^*)^T\boldsymbol{\Sigma}\mathbf{w}^* = \lambda_1$

9.1 PCA with arbitrary k, $\mathcal{L}^{(k)} = \sum_{i=k+1}^d \lambda_i$

$(\mathbf{W}^*, \mathbf{z}_1^*, ..., \mathbf{z}_n^*) = \arg\min_{\mathbf{W}\in\mathbb{R}^{d\times k}:\mathbf{W}^T\mathbf{W}=\mathbf{I}_k,\ \mathbf{z}\in\mathbb{R}^k} \sum_{i=1}^n ||\mathbf{W}^T\mathbf{z}_i - \mathbf{x}_i||_2^2$

Sol.: $\mathbf{W}^* = (\mathbf{v}_1|...|\mathbf{v}_k)$, $\mathbf{z}_i^* = \mathbf{W}^{*T}\mathbf{x}_i$, $\mathbf{x}_{new} = \mathbf{W}^*\mathbf{z}_i^*$

9.2 Connection to SVD

$n\boldsymbol{\Sigma} = \mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{S}^T\mathbf{U}^T\mathbf{U}\mathbf{S}\mathbf{V}^T = \mathbf{V}^T\Lambda\mathbf{V} \quad \rightarrow \lambda_i = \sigma_i^2/n$

Top k principal components are first k columns of $\mathbf{V}$

9.3 Kernel PCA

$k=1$: $\alpha* = \arg\max_{\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}=1} \boldsymbol{\alpha}^T\mathbf{K}^T\mathbf{K}\boldsymbol{\alpha} = \frac{1}{\sqrt{\lambda_1}}\mathbf{v}_1 \quad z^* = a^*k(x)$

$z_k^{(i)} = \sum_{j=1}^n \alpha_j^{(i)} k(\mathbf{x}_j, \mathbf{x}_k)$, proj. of $\mathbf{x}_k$ onto ith feature.

9.4 Autoencoders

If act. f is I, then fitt. a NN autoenc. is equiv. to PCA.

10 Probabilistic Modeling & Interference

$\underbrace{p(\theta|\mathcal{D})}_{\text{posteriorbelief}} = \underbrace{\frac{p(\mathcal{D}|\theta)}{p(\mathcal{D})}}_{\text{update}} \underbrace{p(\theta)}_{\text{priorbelief}} \qquad D_{\mathrm{KL}}(P\,||\,Q) \geq 0$

10.1 MLE

$\widehat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta\in\Theta} p(\mathcal{D};\theta) = \arg\max_{\theta\in\Theta} \prod_{i=1}^n p(\mathbf{x}_i, y_i;\theta)$

$= \arg\min_{\theta\in\Theta} \sum_{i=1}^n -\log p(\mathbf{x}_i, y_i;\theta)$

Kullbach-Leibler: $D_{\mathrm{KL}}(P\,||\,Q) = \mathbb{E}_{X\sim P}[log\frac{p(X)}{q(X)}]$

10.2 MAP

$\widehat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta\in\Theta} p(\theta|\mathcal{D}) = \arg\max_{\theta\in\Theta} p(\mathcal{D}|\theta)p(\theta)$

$= \arg\max_{\theta\in\Theta} \left(\prod_{i=1}^n p(\mathbf{x}_i, y_i|\theta)\right) \cdot p(\theta)$

$= \arg\min_{\theta\in\Theta} \sum_{i=1}^n -\log p(\mathbf{x}_i, y_i|\theta) - \log p(\theta)$

If prior $p(\theta) = 1 \quad \rightarrow \quad \widehat{\theta}_{\mathrm{MLE}} = \widehat{\theta}_{\mathrm{MAP}}$

10.3 Probabilistic Perspective on Regression

Gaussian dist. for data noise: $\widehat{\theta}_{\mathrm{MLE}}$ = min of square loss

Prior for θ: $\widehat{\theta}_{\mathrm{MAP}}$ = solution for ridge/LASSO regression

Gaussian Prior: $\lambda = \frac{\sigma^2}{\sigma_\theta^2}$ Laplacian Prior: $\lambda = \frac{2\sigma^2}{b}$

10.4 Probabilistic Perspective on Classification

Bernoulli distr. for data noise: $\widehat{\theta}_{\mathrm{MLE}}$ = min of log. loss

Prior for θ: $\widehat{\theta}_{\mathrm{MAP}}$ = sol. for log. ridge/LASSO regression

Gaussian Prior: $\lambda = \frac{1}{2\sigma_\theta^2}$ Laplacian Prior: $\lambda = \frac{1}{b}$

10.5 Gaussian Bayes Classifiers (Supervised)

Gaussian Naive Bayes Model: $\Sigma_y = diag(\sigma_1, ..., \sigma_d)$

LDA: $\Sigma_i = \Sigma_j$, Fisher's LDA: $\Sigma_i = \Sigma_j$ & $P(Y=y) = \frac{1}{2}$

10.6 Gaussian Mixture Models (GMM) (Uns.)

Sample length of multivariate Gaussian: $\sigma\sqrt{d}$

GMM with identical, spherical cov. matrix → K-means

Hard EM Algorithm (maybe for Gauss!):

E-step: $z_i^{(t)} = \arg\max_z \mathbb{P}(z|\theta^{(t-1)})\mathbb{P}(\mathbf{x}_i|z, \theta^{(t-1)})$

M-step: $\theta^{(t)} = \arg\max_\theta \mathbb{P}(D^{(t)}|\theta)$

Soft EM Algorithm (needed?):

Constrained GMMs:

$\hat{\Sigma}_y = \frac{1}{n_y}\sum_{i:y_i=y}(\mathbf{x}_i - \widehat{\mu}_y)(\mathbf{x}_i - \widehat{\mu}_y)^T$

11 Tricks

$\nabla_\mathbf{w}\langle\mathbf{w}, \mathbf{x}\rangle = \mathbf{x}$ Proj. Matrix: $\mathbf{P}^2 = \mathbf{P}$ → for ex.: $\mathbf{W}\mathbf{W}^T$

$\nabla_{\mathbf{w}_2}(\mathbf{W}_2\mathbf{W}_1\mathbf{x}) = \mathbf{x}^T\mathbf{W}_1^T \qquad \nabla_{\mathbf{w}_1}(\mathbf{W}_2\mathbf{W}_1\mathbf{x}) = \mathbf{W}_2^T\mathbf{x}^T$

$\mathbb{E}[Y|X=x] = \sum_a a\,\mathbb{P}(Y=a|X=x)$

$\mathbb{E}[\ell(f(X), Y)|X=\mathbf{x}] = \sum_{a=1}^K \ell(f(\mathbf{x}), a)\mathbb{P}(Y=a|X=\mathbf{x})$

$\mathbb{P}(Y \neq a|X=x) = 1 - \mathbb{P}(Y=a|X=x)$

$\mathbb{P}(X) = \sum_i \mathbb{P}(X|Y_i)\mathbb{P}(Y_i)$

PCA loss $L^{(r)} = 0$, if data lives in $\leq$ r-dim. subs. of $\mathbb{R}^d$

$L^{(k)} = \sum_{i=k+1}^d \lambda_i \qquad \rightarrow \qquad \lambda_k = L^{(k-1)} - L^{(k)}$

$||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 = ||\mathbf{y}||_2^2 - 2\mathbf{y}^T\mathbf{X}\mathbf{w} + ||\mathbf{X}\mathbf{w}||_2^2$

Linear Regression: $\mathbb{E}_\epsilon[\mathbf{x}_{\text{test}}^T\widehat{\mathbf{w}}] = \mathbb{E}_\epsilon[\mathbf{X}_{\text{test}}^T((\mathbf{X}^T\mathbf{x})^{-1}\mathbf{X}^T\mathbf{y})]$

$= \mathbb{E}_\epsilon[\mathbf{x}_{\text{test}}^T((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{w}^* + \boldsymbol{\epsilon}))]$

$= \mathbf{x}_{\text{test}}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w}^* + \mathbf{x}_{\text{test}}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}_\epsilon[\boldsymbol{\epsilon}]$

$= \mathbf{x}_{\text{test}}^T\mathbf{w}^*$

K-means: memory, add. and mult. → ploy. in n, d, K

Uniform: $\mathcal{Y} \sim \mathcal{U}[a,b] \rightarrow p(y) = \begin{cases} \frac{1}{b-a}, & y \in [a,b], \\ 0, & else. \end{cases}$

Gaussian: $\mathcal{Y} \sim \mathcal{N}(\mu, \sigma^2) \rightarrow p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$

Laplacian: $\mathcal{Y} \sim \text{Laplace}(\mu, b) \rightarrow p(y) = \frac{1}{2b} e^{-\frac{|y-\mu|}{b}}$

Exponential: $\mathcal{Y} \sim \text{Exp}(\lambda) \rightarrow p(y) = \begin{cases} \lambda e^{-\lambda y}, & y \geq 0, \\ 0, & y < 0. \end{cases}$

Binomial: $\mathcal{Y} \sim \text{Binom}(q, n) \rightarrow p(k) = \binom{n}{k} q^k (1-q)^{n-k}$

Beta dist.: $\mathcal{Y} \sim \text{Beta}(\alpha, \beta) \rightarrow p(y) = c \cdot y^{1-\alpha}(1-y)^{1-\beta}$

Pareto: $\mathcal{Y} \sim \text{Pa}(\mu, k) \rightarrow p(y) = \begin{cases} \frac{k\mu^k}{y^{k+1}}, & y \geq \mu, \\ 0, & y < \mu. \end{cases}$

Bernoulli: $\mathcal{Y} \sim \text{Ber}(p) \rightarrow p(y) = \begin{cases} p, & y = 1, \\ 1-p, & y = -1. \end{cases}$

$\text{Ber}(\sigma(z)) \rightarrow p(y|z) = \begin{cases} \sigma(z), & y = 1, \\ 1-\sigma(z), & y = -1. \end{cases} = \sigma(yz)$

Logistic f: $\sigma(z) = \frac{1}{1+e^{-z}}$, Batch norm: $\overline{x}_i = \gamma\frac{x_i - \mu_S}{\sigma_S} + \beta$

a polynomial of degree m can interp. m+1 points

Orthon. sett.: coord. of $\mathbf{w}_{\ell_1}$ to 0, remain 0 for bigger λ

$\text{trace}(\mathbf{ABC}) = \text{trace}(\mathbf{CAB}) \qquad \text{trace}(\mathbf{A}) = \sum_i \lambda_i$