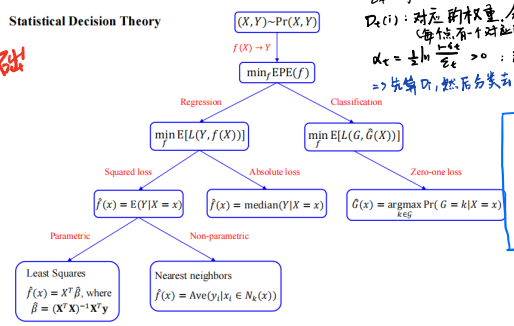
CS182 CheatSheet

Overview to Superviced Learninng

Statistical Decision Theory



- Squared error loss: $L(Y,f(X))=(Y-f(X))^2$
- Expected Prediction Error (EPE): $EPE(f)=E(Y-f(X))^2=\int(Y-f(x))^2Pr(dx,dy)$
  - Noticed that $Pr(dx,dy)=Pr(Y|X)Pr(X)$, we have $EPE(f)=E_XE_Y([Y-f(x)]^2|X)$

Local Methods in High Dimensions

Curse of Dimensionality (维度的诅咒)

1. Local neighborhoods become **increasingly global**, as the number of dimension increases
2. In high dimensions, all samples are **close to the edge** of the sample
3. Samples **sparsely populate** the input space

The Bias-Variance Decomposition:

1. Deterministic Case: $\mathrm{MSE}(x_0)=\mathrm{Var}(\hat{y}_0)+\mathrm{Bias}^2(\hat{y}_0)$

$$\begin{aligned}\mathrm{MSE}(x_0)&=E[f(x_0)-\hat{y}_0]^2\\&=E[\hat{y}_0-E(\hat{y}_0)+E(\hat{y}_0)-f(x_0)]^2\\&=E[(\hat{y}_0-E(\hat{y}_0))^2+2(\hat{y}_0-E(\hat{y}_0))(E(\hat{y}_0)-f(x_0))+(E(\hat{y}_0)-f(x_0))^2]\\&=E[(\hat{y}_0-E(\hat{y}_0))^2]+(E(\hat{y}_0)-f(x_0))^2\\&=\mathrm{Var}(\hat{y}_0)+\mathrm{Bias}^2(\hat{y}_0)\end{aligned}$$

2. Non-deterministic Case

$$\begin{aligned}\mathrm{EPE}(x_0)&=\mathrm{MSE}(x_0)+\sigma^2\\&=\mathrm{Var}(\hat{y}_0)+\mathrm{Bias}^2(\hat{y}_0)+\sigma^2\end{aligned}$$

Linear Regression 线性回归

Simple-Multiple Linear Regression

For $\hat{\mathbf{y}}=\mathbf{X}\beta$, Minimize $\mathrm{RSS}(\beta)=(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)$

- $\frac{\partial \mathrm{RSS}}{\partial\beta}=-2\mathbf{X}^T(\mathbf{y}-\mathbf{X}\beta)=0\Rightarrow\mathbf{X}^T(\mathbf{y}-\mathbf{X}\beta)=0$

So we have $\hat{\beta}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

Gauss-Markov Theorem

The least squares estimator has the lowest sampling variance within the class of linear unbiased estimators.

> 最小二乘估计器在线性无偏估计器类中具有最低的抽样方差。

Proof: suppose $\tilde{\beta}=\mathbf{C}\mathbf{y}$ is a linear estimator of $\hat{\beta}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, where $\mathbf{C}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T+\mathbf{D}$, and $\mathbf{D}\in\mathbb{R}^{p\times N}$ is a non-zero matrix

$$\begin{aligned}\mathrm{E}[\tilde{\beta}]&=\mathrm{E}[C\mathbf{y}]\\&=\mathrm{E}[((X^TX)^{-1}X^T+D)(X\beta+\varepsilon)]\\&=((X^TX)^{-1}X^T+D)X\beta+((X^TX)^{-1}X^T+D)\underbrace{\mathrm{E}[\varepsilon]}_{E[\varepsilon]=0}\\&=((X^TX)^{-1}X^T+D)X\beta\\&=(X^TX)^{-1}X^TX\beta+DX\beta\\&=(I_p+\underbrace{DX}_{0})\beta.\end{aligned}$$

If and only if $\mathbf{DX}=0$, $\tilde{\beta}$ is unbiased.

$$\begin{aligned}\mathrm{Var}(\tilde{\beta})&=\mathrm{Var}(C\mathbf{y})\\&=C\mathrm{Var}(\mathbf{y})C^T\\&=\sigma^2CC^T\\&=\sigma^2((X^TX)^{-1}X^T+D)(X(X^TX)^{-1}+D^T)\\&=\sigma^2((X^TX)^{-1}X^TX(X^TX)^{-1}+(X^TX)^{-1}X^TD^T+DX(X^TX)^{-1}+DD^T)\\&=\sigma^2(X^TX)^{-1}+\sigma^2(X^TX)^{-1}\underbrace{(DX)^T}_{0}+\sigma^2\underbrace{DX}_{0}(X^TX)^{-1}+\sigma^2DD^T\\&=\sigma^2(X^TX)^{-1}+\sigma^2DD^T\\&=\underbrace{\mathrm{Var}(\tilde{\beta})}_{\text{Positive semidefinite}}+\underbrace{\sigma^2DD^T}_{\mathbf{DX}=0}\end{aligned}$$

$\mathrm{Var}(\beta)=\sigma^2(X^TX)^{-1}$

$\mathrm{Var}(\mathbf{y})=E[\mathbf{y}-E[\mathbf{y}]]^2=\mathrm{Var}(\varepsilon)$

16

Ridge Regression

对 Regression Coefficients 进行收缩（Shrink）：

另一个等价的表述是：

$$\arg\min_{\beta}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{i=1}^{p}x_{ij}\beta_j\right)^2\quad\text{subject to}\quad\sum_{j=1}^{p}\beta_j\le t$$

其中的损失项变为：

求导令之为 0，得到：$\beta^{\mathrm{ridge}}=(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{y}$

Additional insight into ridge regression

- Singular value decomposition (SVD)

$\mathbf{X}=\mathbf{UDV}^T$, $\mathbf{U}^T\mathbf{U}=\mathbf{I}_p$, $\mathbf{V}^T\mathbf{V}=\mathbf{I}_p$

- $\mathbf{U}\in\mathbb{R}^{N\times p}$: its columns span the column space ($\mathbb{R}^N$) of $\mathbf{X}$
- $\mathbf{V}\in\mathbb{R}^{p\times p}$: its columns span the row space ($\mathbb{R}^p$) of $\mathbf{X}$
- $\mathbf{D}\in\mathbb{R}^{p\times p}$: diagonal matrix $(d_1\ge d_2\ge\cdots\ge d_p\ge 0)$
  - Singular values of $\mathbf{X}$
  - if $\exists d_j=0$, $\mathbf{X}$ is singular

Least squares

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\mathrm{ls}}&=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\\&=\mathbf{U}\mathbf{U}^T\mathbf{y},\\&=\sum_{j=1}^{p}\mathbf{u}_j\mathbf{u}_j^T\mathbf{y}\end{aligned}$$

$\mathbf{u}_j$: The j-th column of $\mathbf{U}$

Ridge regression

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\mathrm{ridge}}&=\mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}\\&=\mathbf{U}\,\mathbf{D}(\mathbf{D}^2+\lambda\mathbf{I})^{-1}\mathbf{D}\,\mathbf{U}^T\mathbf{y}\\&=\sum_{j=1}^{p}\mathbf{u}_j\frac{d_j^2}{d_j^2+\lambda}\mathbf{u}_j^T\mathbf{y},\end{aligned}$$

- shrinkage factor
- smaller d_j leads to a larger shrinkage

The Lasso

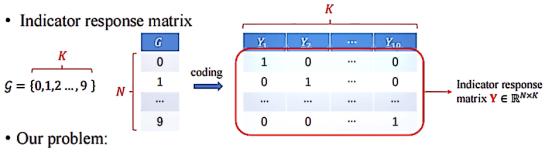
岭回归的 L2-Loss $||\beta||_2=\sum_{j=1}^{p}|\beta_j|^2$ 变为 L1-Loss $||\beta||_1=\sum_{j=1}^{p}|\beta_j|$

Linear Classification 线性分类

Linear regression of an indicator matrix

Categorical output variable G with values from $\mathcal{G}=\{1,\dots,K\}$.

- Indicator response matrix



$\mathcal{G}=\{0,1,2\dots,9\}$

- Our problem:

$$\hat{\mathbf{B}}=\underset{\mathbf{B}}{\mathrm{argmin}}\|\mathbf{Y}-\mathbf{X}\mathbf{B}\|_F^2\qquad \mathbf{B}=(\beta_1,\beta_2,\dots,\beta_{10})\in\mathbb{R}^{(p+1)\times K}$$

- The fitted values on $\mathbf{X}$:

$$\hat{\mathbf{Y}}=\mathbf{X}\hat{\mathbf{B}}=\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}=\mathbf{H}\mathbf{Y}$$

- The zero-one loss $L(k,l)=\begin{cases}1, & k\ne l\\0, & k=l\end{cases}$
- EPE with $Pr(G,X)$: $\mathrm{EPE}=E[L(G,\hat{G}(X))]$
- Pointwise minimize leads to:

$$\hat{G}(x)=\underset{k\in\mathcal{G}}{\arg\min}\sum_{l=1}^{K}L(K,l)Pr(G=l|X=x)=\underset{k\in\mathcal{G}}{\arg\max}\,Pr(G=k|X=x)$$

Basic Idea

Model the posterior $Pr(G=k|X=x)$ based on Bayes Theorem

The Posterior:

$$Pr(G=k|X=x)=\frac{Pr(X=x|G=k)Pr(G=k)}{Pr(X=x)}=\frac{Pr(X=x|G=k)Pr(G=k)}{\sum_{l=1}^{K}Pr(X=x|G=l)Pr(G=l)}$$

其中，$G=k$ 时候 X 的密度为 $f_k(x)=Pr(X=x|G=k)$，而先验有 $\pi_k=Pr(G=k)$

$$Pr(G=k|X=x)=\frac{f_k(x)\pi_k}{\sum_{l=1}^{K}f_l(x)\pi_l}$$

It produces LDA, QDA (quadratic DA), MDA (mixture DA), kernel DA and naive Bayes, under various assumptions on $f_k(x)$

LDA: Linear Discriminant Analysis

Assumptions in LDA

- Model each class density as **Multivariate Gaussian**

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_l^{-1}(x-\mu_k)\right)$$

- Assume the classes share a common covariance $\Sigma_k = \Sigma, \forall k$

Compare two classes k and l

$$\log\frac{Pr(G=k|X=x)}{Pr(G=l|X=x)} = \log\frac{f_k(x)}{f_l(x)} + \log\frac{\pi_k}{\pi_l}$$
$$= \log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k+\mu_l)^T\Sigma^{-1}(\mu_k-\mu_l) + x^T\Sigma^{-1}(\mu_k-\mu_l)$$

二次项由于共同的协方差而消失

- Parameter estimation

$\hat{\pi}_k = N_k/N$, where N_k is the number of class-k observations;

$\hat{\mu}_k = \sum_{g_i=k} x_i/N_k$;

$\hat{\Sigma} = \sum_{k=1}^{K}\sum_{g_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T/(N-K)$.

Pooled covariance (合并方差)

$$\hat{\Sigma} = \frac{(N_1-1)\hat{\Sigma}_1 + (N_2-1)\hat{\Sigma}_2 + \cdots + (N_K-1)\hat{\Sigma}_K}{(N_1-1)+(N_2-1)+\cdots+(N_K-1)}, \text{ where } \hat{\Sigma}_k = \frac{\sum_{g_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T}{N_k-1}$$

Weighted average

QDA: Quadratic Discriminant Analysis

Assume that each class has a specific covariance Σ_k

Discriminant Functions:

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$$

Probability and Estimator

MLE: Maximum Likelihood Estimate

Choose θ that maximizes probability of observed data $\mathcal{D}$

$$\hat{\theta} = \arg\max_\theta P(\mathcal{D}|\theta)$$

MAP: Maximum a Posterior

Choose θ that is most probable given prior probability and the data

Naive Bayes

- Train Naive Bayes:
  - for each value y_k
    - estimate $\pi_k \equiv P(Y=y_k)$
    - for each value x_{ij} of each attribute X_i
      - estimate $\theta_{ijk} \equiv P(X_i = x_{ij}|Y=y_k)$
- Classify
  - $Y^{new} \leftarrow \arg\max_{y_k} P(Y=y_k)\prod_i P(X_i^{new}|Y=y_k)$
  - $Y^{new} \leftarrow \arg\max_{y_k} \pi_k \prod_i \theta_{ijk}$

Estimate Parameters

Maximum likelihood estimates:

$$\hat{\pi}_k = \hat{P}(Y=y_k) = \frac{\#D\{Y=y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i=x_j|Y=y_k) = \frac{\#D\{X_i=x_j \wedge Y=y_k\}}{\#D\{Y=y_k\}}$$

MAP estimates (Beta, Dirichlet priors):

Only difference: "imaginary" examples

$$\hat{\pi}_k = \hat{P}(Y=y_k) = \frac{\#D\{Y=y_k\}+(\beta_k-1)}{|D|+\sum_m(\beta_m-1)}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i=x_j|Y=y_k) = \frac{\#D\{X_i=x_j \wedge Y=y_k\}+(\beta_k-1)}{\#D\{Y=y_k\}+\sum_m(\beta_m-1)}$$

Continuous X_i (Gaussian Naive Bayes)

Assume:

$$P(X_i=x|Y=y_k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}}\exp\left(\frac{-(x-\mu_{ik})^2}{2\sigma_{ik}^2}\right)$$

Maximum likelihood estimates:

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j=y_k)}\sum_j X_i^j\delta(Y^j=y_k)$$

$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j=y_k)}\sum_j (X_i^j-\hat{\mu}_{ik})^2\delta(Y^j=y_k)$$

ith feature; kth class; jth training example; δ(z)=1 if z true, else 0

Kernel Method & SVM

The online Learning Model: Perceptron

Algorithm

- Set $t=1$, start with all zero vector w_1
- Given example x, predict positive iff $w_t \cdot x \geq 0$
- On a mistake, update:
  - Mistake on positive, then update $w_{t+1} \leftarrow w_t + x$
  - Mistake on negative, then update $w_{t+1} \leftarrow w_t - x$
- Note: $w_t = a_{i1}x_{i1} + \cdots + a_{ik}x_{ik}$

Geometric Margin

- Definition: The **margin** of example x w.r.t. a linear sep. w is the distance from x to the plane $w \cdot x = 0$
- The **margin** γ_w of a set of examples S w.r.t. a linear separator w is the smallest margin over points $x \in S$.
- The margin γ of a examples S is the **maximum** γ_w over all linear seps w

Mistake Bound

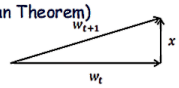
Theorem: If data linearly separable by margin γ and points inside a ball of radius R, then Perceptron makes $\leq R/\gamma^2$ mistakes.

Proof:

Idea: analyze $w_t \cdot w^*$ and $\|w_t\|$, where w^* is the max-margin sep, $\|w^*\| = 1$.

Claim 1: $w_{t+1} \cdot w^* \geq w_t \cdot w^* + \gamma$. (because $l(x)x \cdot w^* \geq \gamma$)

Claim 2: $\|w_{t+1}\|^2 \leq \|w_t\|^2 + R^2$. (by Pythagorean Theorem)



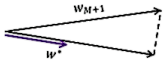
After M mistakes:

$w_{M+1} \cdot w^* \geq \gamma M$ (by Claim 1)

$\|w_{M+1}\| \leq R\sqrt{M}$ (by Claim 2)

$w_{M+1} \cdot w^* \leq \|w_{M+1}\|$ (since w^* is unit length)

So, $\gamma M \leq R\sqrt{M}$, so $M \leq \left(\frac{R}{\gamma}\right)^2$.



SVMs

Directly optimize for the maximum margin separator:

- Input: $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$
- Maximize γ under:
  - $||w||^2 = 1$
  - $\forall i, y_i w \cdot x_i \geq \gamma$

让 $w' = \frac{w}{\gamma}$，把优化问题转化成凸问题：

- Maximize $||w'||^2$ under $\forall i, y_i w' \cdot x_i \geq 1$
- If data isn't perfectly linearly separable?
  - Replace "# mistakes" with upper bound called "hinge loss"
  - Minimize $||w||^2 + C\sum_i \xi_i$ s.t. $\forall i, y_i w \cdot x_i \geq 1 - \xi_i, \xi_i \geq 0$

Lagrangian Dual of SVMs

Solution

We first formulate the Lagrangian function of the primal problem:

$$L(\mathbf{w}, \xi, \alpha, \lambda) = ||\mathbf{w}||_2^2 - \sum_{i=1}^{n} \alpha_i (y_i \mathbf{w}^\top x_i - 1),$$

where $\alpha_i \geq 0$ $(\forall i)$ is the dual variable. Because strong duality holds in the primal problem, the optimal optimization variables $\{\mathbf{w}^*, \alpha^*\}$ should satisfy K.K.T. conditions:

- primal: $y_i \mathbf{w}^{*\top} x_i \geq 1, \forall i,$
- dual: $\alpha_i^* \geq 0, \forall i,$
- complementary: $\alpha_i^* (y_i \mathbf{w}^{*\top} x_i - 1) = 0, \forall i,$
- stationary: $\nabla_{\mathbf{w}^*} L = 0.$

According to the stationary condition, we have

$$\nabla_{\mathbf{w}} L = 2\mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i x_i = 0, \quad \Rightarrow \quad \mathbf{w} = \frac{1}{2} \sum_{i=1}^{n} \alpha_i y_i x_i,$$

Substituting them into the Lagrangian function yields the dual function $g(\alpha, \lambda)$,

$$\begin{aligned} g(\alpha) &= \inf_{\mathbf{w}, \xi} L(\mathbf{w}, \alpha) \\ &= \frac{1}{4} (\sum_{i=1}^{n} \alpha_i y_i x_i, \sum_{j=1}^{n} \alpha_j y_j x_j) - \sum_{i=1}^{n} \alpha_i (y_i (\frac{1}{2} \sum_{j=1}^{n} \alpha_j y_j x_j, x_i) - 1) \\ &= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i. \end{aligned}$$

Thus, the dual problem is

$$\begin{aligned} \max_{\alpha} \quad & -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle + \sum_{i=1}^{n} \alpha_i, \\ \text{s.t.} \quad & \alpha_i \geq 0, \ \forall i. \end{aligned}$$

Kernels

$K(\cdot, \cdot)$ is a kernel if it can be viewed as a legal definition of inner product:

$$\exists \phi : X \to \mathbb{R}^N, \text{ s.t. } K(x, z) = \phi(x) \cdot \phi(z)$$

Theorem (Mercer)

K is a kernel if and only if:

- K is symmetric
- For any set of training points $x_1, x_2, \ldots, x_m$ and for any $a_1, a_2, \ldots, a_m \in \mathbb{R}$, we have:

$$\sum_{i,j} a_i a_j K(x_i, x_j) \geq 0$$

$$a^T K a \geq 0$$

I.e., $K = (K(x_i, x_j))_{i,j=1,\ldots,n}$ is positive semi-definite

Kernels

- Linear: $K(x, z) = x \cdot z$
- Polynomial: $K(x, z) = (x \cdot z)^2$ or $K(x, z) = (1 + x \cdot z)^d$
- Gaussian: $K(x, z) = \exp\left[-\frac{||x-z||^2}{2\sigma^2}\right]$
- Laplace: $K(x, z) = \exp\left[-\frac{||x-z||}{2\sigma^2}\right]$

MLP 全连接 NN

反向传播四公式

BP1

If $Y = \sigma(X)$, then:

$$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial Y} \odot \sigma'(x)$$

Where $\odot$ means element-wise product

BP2 & 3 & 4

If $Y = W \cdot X + B$

$$\frac{\partial L}{\partial X} = W^T \cdot \frac{\partial L}{\partial Y} \qquad \frac{\partial L}{\partial B} = \frac{\partial L}{\partial Y} \qquad \frac{\partial L}{\partial W} = \frac{\partial L}{\partial Y} \cdot X^T$$

推导的核心

考虑： $y_i = w_{i1}x_1 + w_{i2}x_2 + \cdots + w_{in}x_n + b_i$

$$\frac{\partial y_i}{\partial w_{ij}} = x_j, \quad \frac{\partial y_i}{\partial x_j} = w_{ij}, \quad \frac{\partial y_i}{\partial b_i} = 1$$

常用损失函数的导数

Softmax

If $X = [x_1, \cdots, x_n], Y = \text{SoftMax}(X) = [y_1, \cdots, y_n]$

Thus $y_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_i}}$ and obviously $\sum_{i=1}^{n} y_i = 1$

Thus we have:

$$\frac{\partial y_i}{\partial x_j} = \begin{cases} y_i(1 - y_i), & i = j \\ -y_i y_j, & u \neq j \end{cases}$$

It's equal to:

$$\frac{\partial Y}{\partial X} = \text{diag}(Y) - Y^T Y$$

Sigmoid

We have $\sigma(x) = \text{sigmoid}(x) = \frac{1}{1+e^{-x}}$

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

Tanh

We have $\tau(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

$$\tau'(x) = 1 - \tau^2(x)$$

Softmax with 交叉熵：

$$L = -\sum_k \hat{y}_k \log y_k$$

We have:

$$\frac{\partial L}{\partial X} = Y - \hat{Y}$$

Dimension Reduction: PCA

Denoted that $v_1, \ldots, v_d$ are the d principal components, $v_i \cdot v_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$

Let $X = [x_1, \ldots, x_n]$ (Columns are the datapoints)

Maximizes sample variance of projected data

$$\frac{1}{n} \sum_{i=1}^{n} (\mathbf{v}^T x_i)^2 = \mathbf{v}^T \mathbf{X}\mathbf{X}^T \mathbf{v} \text{ s.t. } \mathbf{v}^T \mathbf{v} = 1$$

Lagrangian: $\arg\max_{\mathbf{v}} \mathbf{v}^T \mathbf{X}\mathbf{X}^T \mathbf{v} - \lambda \mathbf{v}^T \mathbf{v}$

We finally have $(\mathbf{X}\mathbf{X}^T - \lambda \mathbf{I})\mathbf{v} = 0, (\mathbf{X}\mathbf{X}^T)\mathbf{v} = \lambda \mathbf{v}$

Maximum Variance Direction: 1st PC a vector v such that projection on to this vector capture maximum variance in the data (out of all possible one dimensional projections)

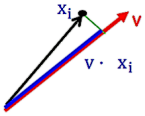
$$\frac{1}{n} \sum_{i=1}^{n} (\mathbf{v}^T \mathbf{x}_i)^2 = \mathbf{v}^T \mathbf{X}\mathbf{X}^T \mathbf{v} \qquad \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{x}_i - (\mathbf{v}^T \mathbf{x}_i)\mathbf{v}\|^2$$

Minimum Reconstruction Error: 1st PC a vector v such that projection on to this vector yields minimum MSE reconstruction

blue² + green² = black²

black² is fixed (it's just the data)

So, maximizing blue² is equivalent to minimizing green²



> 这里的 $\mathbf{v}^T\mathbf{X}$ 是投影，再乘上 $\mathbf{v}$ 就是 在投影上的向量。

- The eigenvalue λ denotes the amount of variability captured along that dimension (aka amount of energy along that dimension).
- Zero eigenvalues indicate no variability along those directions $\Rightarrow$ data lies exactly on a linear subspace
- Only keep data projections onto principal components with non-zero eigenvalues, say $v_1, \ldots, v_k$, where $k = \text{rank}(\mathbf{X}\mathbf{X}^T)$

Clustering: KMeans

Denotions

- Given a sample $\mathcal{X} = \{\mathbf{x}^t\}_{t=1}^N$
- Find k reference vectors $\mathbf{m}_j$ which best represent the data

Encoding/Decoding

Each data point $\mathbf{x}^t$ is represented by the index i of the nearest reference vector $i = \arg\min_j ||\mathbf{x}^t - \mathbf{m}_j||$

We can use labels $\mathbf{b}^t$ for $\mathbf{x}^t$ as:

$$b_i^t = \begin{cases} 1 & \text{if } i = \arg\min_j ||\mathbf{x}^t - \mathbf{m}_j|| \\ 0 & \text{otherwise} \end{cases}$$

The Total Reconstruction Error:

$$E(\{\mathbf{n}_i\}_{i=1}^k | \mathcal{X}) = \sum_t \sum_i b_i^t ||\mathbf{x}^t - \mathbf{m}_i||^t$$

Optimization

$$\arg\min_{\{\mathbf{m}_i\}_{i=1}^k, \{\mathbf{b}^t\}_{t=1}^N} \sum_t \sum_i b_i^t ||\mathbf{x}^t - \mathbf{m}_i||^2$$

However, since b_i also depends on $\mathbf{m}_i$, the optimization problem cannot be solved analytically, but **iteratively**

Algorithm

Algorithm

Initialize $\mathbf{m}_i$, $i = 1, \ldots, k$ (e.g., k randomly selected $\mathbf{x}^t$)
Repeat
For all $\mathbf{x}^t \in \mathcal{X}$, we obtain the estimated labels

$$b_i^t = \begin{cases} 1 & \text{if } i = \arg\min_j \|\mathbf{x}^t - \mathbf{m}_j\| \\ 0 & \text{otherwise} \end{cases}$$

For all $\mathbf{m}_i$, $i = 1, \ldots, k$, we obtain (by taking the derivative of $E(\{\mathbf{m}_i\}_{i=1}^k \mid \mathcal{X})$ with respect to $\mathbf{m}_i$ and setting it to $\mathbf{0}$)

$$\mathbf{m}_i = \frac{\sum_t b_i^t \mathbf{x}^t}{\sum_t b_i^t}$$

The reference vector is set to the mean (center) of all the instances that it represents.
Until $\mathbf{m}_i$ converge.

- Note $\mathbf{m}_i$ here is the same as the formula for the mean estimation in classification, except that we place the estimated labels $\mathbf{b}^t$ in place of the labels $\mathbf{r}^t$.

梯度下降

Gradient Descent

CNN 参数计算. 输入为 W×H×C, 卷积核为 F×F×C, 为 K个. Padding 为 P, 步长 stride = S, 为有 L 层

Basic Problem

$$\arg\min_{x \in \mathbb{R}^n} f(x)$$

每层大小: $W' = \frac{W-F+2P}{S}+1$ $H' = \frac{H-F+2P}{S}+1$ $\Rightarrow W' \times H' \times K$

Iteration: $x^{r+1} = x^r - \gamma_r \cdot \nabla f(x^r)$

每层参数: $(F \times F \times D + 1) \times K$

全连接层: $(\text{Feature\_in} + 1) \times \text{Feature\_out}$.

Convexity

$$f(\lambda(x) + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$$
$$f(x) \ge f(y) + \nabla f(y)^T (x - y)$$
$$\nabla^2 f(x) \ge 0$$

L-smooth

$$||\nabla f(x) - \nabla f(y)|| \le L||x - y||$$

Descemdent Lemma

$$|f(x) - f(y) - \nabla f(y)^T (x - y)| \le \frac{L}{2}||x - y||^2$$

If f twice-differentiable, L-smooth $\Leftrightarrow \nabla^2 f(x) \le LI$, $\quad \mathbf{d}^T \nabla^2 f(x) \mathbf{d} \le L||\mathbf{d}||^2, \forall \mathbf{d}$

Convergence Analysis

Optimality measure $M(x^r)$

- Convex: $||x^r - x^*||$ or $f(x^r) - f^*$
- Non-convex: $||\nabla f(x^r)||$

Order of convergence q s.t.

$$\sup\left\{q \middle| \lim_{r \to +\infty} \frac{M(x^{r+1})}{M(x^r)^q} < \infty\right\}$$

- $q = 1$: Linear convergence, $q = 2$: quadratic

Rate of Convergence: Given q ,

$$\lim_{r \to \infty} \frac{M(r^{r+1})}{M(x^r)^1} = n$$

- Sublinear: $\text{Rate} = 1$, Superlinear: $\text{Rate} = 0$

Convergence under Convexity

Theorem
Let f be convex with bounded gradient $\|\nabla f(\mathbf{x}^r)\| \le B$, then the sequence $(\mathbf{x}^r)_{r \in \mathbb{N}}$ generated by GD with step size $\gamma_r = \frac{f(\mathbf{x}^r) - f^*}{\|\nabla f(\mathbf{x}^r)\|^2}$ satisfies

$$\min_{r=0,\ldots,T-1} f(\mathbf{x}^r) - f^* \le \frac{B\|\mathbf{x}^0 - \mathbf{x}^*\|}{\sqrt{T}}.$$

Convergence under Smoothness

Theorem
Let f be L-smooth, then the sequence $(\mathbf{x}^r)_{r \in \mathbb{N}}$ generated by GD with step size $\gamma \le 1/L$ satisfies

$$\min_{r=0,\ldots,T-1} \|\nabla f(\mathbf{x}^r)\|^2 \le \frac{2}{\gamma} \frac{(f(\mathbf{x}^0) - f(\mathbf{x}^*))}{T}.$$

Convexity & Smoothness

Theorem
Let f be convex and L-smooth, then the sequence $(\mathbf{x}^r)_{r \in \mathbb{N}}$ generated by GD with step size $\gamma \le 1/L$ satisfies

$$f(\mathbf{x}^T) - f^* \le \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|}{2\gamma T}.$$

Upper & Lower Bound

μ-Strong Convexity

$$f(x) \ge f(y) + \nabla f(y)^T (x - y) + \frac{\mu}{2}||x - y||_2^2$$

L-Smoothness

$$f(x) \le f(y) + \nabla f(y)^T (x - y) + \frac{L}{2}||x - y||^2$$

Implication

$$\nabla f(x^r)^T (x^r - x^*) \ge f(x^r) - f_{\frac{r}{2}}^* ||x^r - x^*||$$
$$f(x^{r+1}) \le f(x^r) - \frac{r}{2}||\nabla f(x^r)||^2$$

Lagrangian 拉格朗日

Basics

Original Function

Minimize $f_0(x)$ with optimal p^* subject to $f_i(x) \le 0, i = 1, \ldots, m$
We have $L(x, \lambda) = f_0(x) + \lambda_1 f_1(x) + \cdots + \lambda_m f_m(x)$

Dual Function

Lower Bound Property: if $\lambda \ge 0$ and x are primal feasible:

$$g(\lambda) \le f_0(x)$$

Dual Problem

Maximize $g(\lambda)$ with optimal d^* subject to $\lambda \ge 0$
We have: $d^* \le p^*$, denoted that $p^* - d^*$ the optimal dual gap.
The Convec Problem have $p^* = d^*$

KKT Optimal Condition

$f_i(x^*) \le 0$	Primal Feasible
$\lambda_i^* \ge 0$	Dual Feasible
$\lambda_i^* f_i(x_i^*) = 0$	Complementary
$\nabla f_0(x^*) + \Sigma \lambda_i^* \nabla f_i(x^*) = 0$	Stationary

Equality Constraints

$\partial A = 0$
$\partial(\alpha X) = \alpha \partial X$
$\partial(X+Y) = \partial X + \partial Y$
$\partial(\text{Tr}(X)) = \text{Tr}(\partial X)$
$\partial(XY) = (\partial X)Y + X \partial Y$
$\partial(X^{-1}) = -X^{-1}(\partial X)X^{-1}$
$\partial(\det(X)) = \text{Tr}(\text{adj}(X)\partial X)$
$= \det(X)\text{Tr}(X^{-1}\partial X)$
$\ln \det(X) = \det(X)\text{Tr}(X^{-1}\partial X)$
$\partial(X^T) = (\partial X)^T$

$$\frac{\partial x^T a}{\partial x} = \frac{\partial a^T x}{\partial x} = a$$
$$\frac{\partial a^T X b}{\partial X} = ab^T$$
$$\frac{\partial a^T X^T b}{\partial X} = ba^T$$
$$\frac{\partial a^T X a}{\partial X} = \frac{\partial a^T X^T a}{\partial X} = aa^T$$
$$\frac{\partial x}{\partial x_{ij}} = J^{ij}$$

Minimize $f_0(x)$ subject to $f_i(x) \le 0, i = 1, \ldots, m$ and $h_i(x) = 0, i = 1, \ldots, p$
We have $L(x, \lambda, v) = f_0(x) + \sum \lambda_i f_i(x) + \sum v_i h_i(x)$

- Dual Function: $g(\lambda, v) = \inf_x (L(x, \lambda, v))$
- Dual Problem: Maximize $g(\lambda, v)$ subject to $\lambda \ge 0$

KKT:

$f_i(x^*) \le 0, h_i(x^*) = 0$	Primal Feasible
$\lambda_i^* \ge 0$	Dual Feasible
$\lambda_i^* f_i(x_i^*) = 0$	Complementary
$\nabla f_0(x^*) + \sum \lambda_i^* \nabla f_i(x^*) + \sum v_i^* \nabla h_i(x^*) = 0$	Stationary

LTSM: