PRNN Notes

Lokesh Mohanty

April 25, 2023

Contents

1 Naive Baye's Classifier (02/03/2023)

Here we assume that in the features are independent of each other which makes it is easy to find the class conditional density h

$$h_B(x) = \begin{cases} 1, P_{y=1/x} > P_{y=0/x} \\ 0, otherwise \end{cases}$$
$$p_{x/y=i} = \Pi_{j=1}^{d} p_{xy/y=i}$$

2 The Bias-Variance decomposition

$R(h) = Bias^2 + Variance + Noise^2$

$$\begin{aligned} Bias &= \mathbb{E}[\overline{h}(x) - h^*(x)] \\ Variance &= \mathbb{E}[h_D(x) - \overline{h}(x)]^2 \\ Noise &= \mathbb{E}[h(x) - h^*(x)] \end{aligned}$$

3 Regularization

ERM: $\min_\theta \hat{R}(h_\theta)$ Regularized ERM: $\min_\theta \hat{R}(h_\theta) + \lambda\Omega(\theta)$, (Regularizer) $\Omega(\theta) : \theta \to \mathbb{R}$, (Regularization constant) λ

Claim: A regularized h_θ will have more bias compared to its unregularized counterpart

Eg: $y = ax^2 + \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, $x \in \mathbb{R}$ $D = (x_i, y_i)_{i=1}^{6}$

$$
\begin{aligned}
h_1^w(x_i) &= \sum_{j=1}^{6} w_i x_i^j + w_0 \\
h_1^{w*}(x) &= \arg\min_w \hat{R}(h_1^w) \\
\hat{h}_1^{w*}(x) &= \arg\min_w \left[\hat{R}(h_1^w + \lambda ||w||_2^2)\right] \\
\Omega : w \to \mathbb{R}^+ &= ||w||_2^2 \\
\text{Bias } \hat{h}_1^{w*} &> \text{Bias } h_1^{w*}
\end{aligned}
$$

It can also be considered as a constrained optimization problem (i.e., min $\hat{R}(w)$ s.t. $\Omega(w) \leq k$(some constant))

4 Fisher Linear Discriminant Analysis (FLDA)

$$
h_b(x) = \begin{cases} 0, \; p_{y=0/x} - p_{y=1/x} > 0 \\ 1, \; otherwise \end{cases}
$$

Discriminant function: $P_{y=0/x} - P_{y=1/x}$

4.1 Linear Separability

A given $\mathcal{D}$, is called linearly separable, if there exists a linear discriminant function such that $\forall x_i \in \mathcal{D}$, $w^T x_i > 0$ if $y_i = 0$ and $w^T x_i < 0$ if $y_i = 1$.

Given a $\mathcal{D}$ (linearly separable), find a **good w**. (maximize the difference between the projected means of the different classes) suppose $y_i \in c_1, c_2$, let $\mu_1 = \frac{1}{n_1} \sum_{i, y_i \in c_1} w^T x_i$, $\mu_2 = \frac{1}{n_2} \sum_{i, y_i \in c_2} w^T x_i$, $s_1^2 = \sum_{i: x_i \in c_1} (w^T x_i - \mu_1)^2$, $s_2^2 = \sum_{i: x_i \in c_2} (w^T x_i - \mu_1)^2$

$$
\begin{aligned}
R_{FLDA}(w) &= \frac{\mu_1^2 - \mu_2^2}{s_1^2 + s_2^2} = \frac{W^T S_B W}{W^T S_W W} \\
w^*_{FLDA} &= \arg\max_w R_{FLDA}(w) \to w : S_B W = \lambda S_W W
\end{aligned}
$$

inter-class scatter -> $\mu_1^2 - \mu_2^2$ intra-class scatter -> $s_1^2 + s_2^2$

$$
\begin{aligned}
(\mu_1-\mu_2)^2 &= W^T S_B W \\
S_B^{d\times d} &= (m_1-m_2)(m_1-m_2)^T \rightarrow \text{ between class scatter matrix} \\
S_1^2 &= W^T \left[\sum_{i\in C_1}(x_i-m_1)(x_i-m_1)^T\right] W \\
S_2^2 &= W^T \left[\sum_{i\in C_2}(x_i-m_2)(x_i-m_2)^T\right] W \\
\text{Define } S_w &= \sum_{i\in C_1}(x_i-m_1)(x_i-m_1)^T \rightarrow \text{ with class scatter matrix}
\end{aligned}
$$

Observe,

$$S_B W = (m_1-m_2)(m_1-m_2)^T W = k(m_1-m_2), \text{ (as } (m_1-m_2)^T W \text{ is scalar)}$$

Since, $S_B W = \lambda S_W W \implies k(m_1-m_2) = \lambda S_W W \implies W_{FLDA} = c.S_W^{-1}(m_1-m_2)$

5 Perceptron Training Algorithm (28/02/2023)

Let w_k, x_k and y_k denote w, x and y at the kth iteration Define $\Delta w_k = w_{k+1} - w_k$, such that

$$\Delta w_k = \begin{cases} 0, \text{ if } (w_k^T x_k > 0 \text{ and } y_k = 1) \text{ or } (w_k^T x_k < 0 \text{ and } y_k = 0) \\ x_k, \text{ if } w_k^T x_k \leq 0 \text{ and } y_k = 1 \\ -x_k, \text{ if } w_k^T x_k \geq 0 \text{ and } y_k = 0 \end{cases}$$

5.1 Claim: This algorithm converges in finite number of steps

Proof by contradiction:

- Multiply all x_i with $y_i = 0$ by -1.
- Then, $w^T x_i > 0 \ \forall i$.
- $w_{k+1} = w_k + x_k$ if $w_k^T x_k \leq 0$.

Assume that the algorithm fails to find a seperating hyperplane. Then, $w_k^T x_k \leq 0 \; \forall k$. (counring only misclassifications)

$$\begin{aligned}
w_{k+1} &= w_k + x_k \\
\|w_{k+1}\|^2 &\leq \|w_k\|^2 + \|x_k\|^2 \; [\because w_k^T x_k \leq 0] \\
\|w_k\|^2 &\leq \|w_0\|^2 + \sum_{i=0}^{k-1} \|x_i\|^2 \\
&\text{without the loss of generality, assume } w_0 = [0...0]^{d\times 1} \\
\|w_k\|^2 &\leq \sum_{i=0}^{k-1} \|x_i\|^2 \\
\text{let } M &= \max_i \|x_i\|^2 \\
\implies \|w_k\|^2 &\leq kM
\end{aligned}$$

Since data is linearly separable, $\exists$ a w^* such that $x_i^T w^* > 0, \forall i$ Let $v = \min_i x_i^T w^*$

$$\begin{aligned}
w_k^T w^* &= \left(\sum_{i=0}^{k-1} x_i \right)^T w^* \\
|w_k^T w^*|^2 &\geq k^2 v^2 \\
\|w_k\|^2 \|w^*\|^2 &\geq k^2 v^2 \text{ cauchy schwartz} \\
\|w^*\|^2 kM &\geq k^2 v^2 \\
k &\leq \frac{\|w^*\|^2 M}{v^2}
\end{aligned}$$

6 Non-parametric Density Estimation (02/03/2023)

Estimate $P_x(x_i)$ directly

Suppose P_x is the density to be estimated

Probability of a point x falling in a region $\mathcal{R}$ (on the support of P_x) is given by

$$P = \int_R P_x(x) dx$$

Suppose sample $x_1, ..., x_n \sim i.i.d. P_x$ Probability of k points out of $n \in \mathcal{R}$

ML estimate for $P = k/n$, where k is the number of data points(from $\mathcal{D}$) that $\in R$

If $\mathcal{R}$ is small and has a volume of $\mathcal{V}$, then $P \approx P(x)\mathcal{V}$ Hence, $P(x)\mathcal{V} = k/n \implies P(x) = \frac{k}{n.\mathcal{V}}$

6.1 Parzen window estimate (generalization of histogram idea)

fix $\mathcal{V}$ and count k R_n is a d dimensional hypercube with length h_0

$$V_n = h_n^d$$

Define a window function:

$$\phi(u) = \begin{cases} 1, |u_j| \leq 1/2, j = 1, ..., d \\ 0, otherwise \end{cases}$$

$$\implies k_n = \sum_{i=1}^{n} \phi\left(\frac{x - x_i}{h_n}\right)$$

$$\implies P_n(x) = \frac{\frac{1}{n}\sum_i = 1^n \phi\left(\frac{x-x_i}{h_n}\right)}{h_n^d}$$

Gaussian kernel: $\phi(u) = exp(-\|u - u_0\|^2)$

6.2 k-nearest neighbour estimates

fix k and grow $\mathcal{V}$,

$$P(x) = \frac{k}{n\mathcal{V}}$$

Suppose we place a volume of $\mathcal{V}$ around a point x and capture k samples Let k_i be the number of points with class i

$$k = \sum_i k_i$$

$$\implies P(x, y_i) = \frac{k_i}{n\mathcal{V}}$$

$$\implies P(y_i/x) = \frac{P(x, y_i}{\sum_i P(x, y_i)}$$

$$\implies P(y_i/x) = \frac{\frac{k_i}{n\mathcal{V}}}{\frac{k}{n\mathcal{V}}} = \frac{k_i}{k}$$

6.2.1 knn classifier

$$h_\theta(x) = \begin{cases} 1, k_1 > k_0 \\ 0, k_0 \geq k_1 \end{cases}$$

knn is a bayes classifier with density coming from non-parametric density estimation knn error is upper bounded by twice of minimum error (bayes error)

7 Support Vector Machines (09/03/2023)

Linearly separable data $\exists w$ such that $w^T x_i + b > 0$ if $y_i = 1$ and < 0 if $y_i = -1$

Hyperplane: $w^T x + b = 0$

$\implies \exists \epsilon > 0$ such that

$$\begin{aligned} w^T x_i + b &\geq \epsilon, \text{ if } y_i = 1 \\ &< -\epsilon, \text{ if } y_i = -1 \\ \implies w^T x_i + b &\geq 1, \text{ if } y_i = 1 \\ &< -1, \text{ if } y_i = -1 \\ \implies y_i(w^T x_i + b) &\geq 1 \; \forall i \end{aligned}$$

Which means that there is no data point between the lines $w^T x + b = 1$ and $w^T x + b = -1$ which are parallel to $w^T x + b = 0$. Distance between the lines is $\frac{2}{\|w\|}$

SVM: $\min_w \frac{1}{2} w^T w$ subject to $y_i(w^T x_i + b) \geq 1 \; \forall i$

7.1 Constrained optimization

$\min f(w)$, $w \in \mathbb{R}^d$ subject to $a_j^T w + b_j \leq 0$, $j = 1, ..., r$, $f : \mathbb{R}^d \to \mathbb{R}$, $a_j \in \mathbb{R}^d$, $b_j \in \mathbb{R} - 1$

Define a Lagrangian funciton $L(w, \mu) = f(w) + \sum_{j=1}^{r} \mu(a_j^T w + b_j)$, $\mu \in \mathbb{R}$, $j = 1, ..., r$ -> lag coefficient

KKT (Karush-Kuhn-Tucker) conditions For a convex $f(w)$, any w^* is a global minima, iff w^* is feasible and $\exists \mu_j^*$, $j = 1, ..., r$ such that

1. $\nabla L(w^*, \mu^*) = 0$
2. $\mu_j^* \geq 0 \; \forall j$
3. $\mu_j^*(a_j^T w^T + b_j) = 0 \; \forall j$

7.2 KKT conditions for SVM

$L(w, b, \mu) = \frac{1}{2}w^T w + \sum_{i=1}^{n} \mu_i \left[1 - y_i(w^T x_i + b)\right]$

1. $\nabla_w L = 0 \implies w^* = \sum_{i=1}^{n} \mu^* y_i x_i$, $\nabla_b L = 0 \implies \sum_{i=1}^{n} \mu^* y_i = 0$
2. $\mu_j^* \geq 0 \ \forall j$
3. $\mu_j^*(a_j^T w^T + b_j) = 0 \ \forall j \implies \mu_i^* \left[1 - y_i(w^{*T} x_i + b^*)\right] = 0 \ \forall i$

Define $S = \{x_i : \mu_i > 0\}$, $w^* = \sum_{i \in S} y_i \mu_i x_i$

7.2.1 Duality (14/03/2023)

$L(w, \mu) = f(w) + \sum_{j=1}^{\gamma} \mu_j(a_j^T w + b_j)$

Dual: $q : \mathbb{R}^\gamma \to \mathbb{R}$ $q(\mu) = \inf_\mu L(w, \mu)$ Dual problem: $\max_\mu q(\mu)$ s.t. $\mu_j \geq 0$, $j = 1, ..., \gamma$

Primal-dual relation: If primal has a solution, dual also has a solution $q(\mu^*) = f(w^*)$

w^* is optimal for primal, μ^* is optimal for dual iff

1. w^* is feasible for primal and μ^* is feasible for dual
2. $f(w^*) = L(w^*, \mu^*) = \min_w L(w, \mu^*) = q(\mu^*)$

7.3 Solution

for SVM primal,

$$q(\mu) = \inf_{w,b} \left\{ \frac{1}{2} w^T w + \sum_{i=1}^{n} \mu_i \left[1 - y_i(w^T x_i + b)\right] \right\} \tag{1}$$

if $\sum_i \mu_i y_i \neq 0$ then $q(\mu) = -\infty$

To prevent this, we add a constraint $\sum_i \mu_i y_i = 0$

$$w^* = \arg\inf_w q(\mu) = \sum_{i \in S} \mu_i y_i x_i, \ \{\text{comes from } \nabla L(w^*, \mu) = 0\}$$

Substitute w^*, b^* and $\sum_i \mu_i y_i = 0$ in equation (1)

$$q(\mu) = \frac{1}{2} w^{*T} w^* + \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \mu_i y_i (w^{*T} x_i + b^*)$$
$$w^* = \sum_i \mu_i y_i x_i, \ \sum_i \mu_i y_i = 0$$

$$
\begin{aligned}
\implies q(\mu) &= \frac{1}{2}\left(\sum_i \mu_i y_i x_i\right)^T + \dots \\
&= \sum_{i=1}^{n} \mu_i - \frac{1}{2}\sum_i \sum_j \mu_i y_i \mu_j y_j x_i^T x_j
\end{aligned}
$$

Note: The dual problem only involves the inner products of the data points.

Dual Problem: (Quadratic programming problem with linear constraints)

$$
\begin{aligned}
&\max_{\mu} \sum_i \mu_i - \frac{1}{2}\sum_i \sum_j \mu_i \mu_j y_i y_j x_i^T x_j \\
&\text{s.t. } \mu_i \geq 0 \; i = 1 \ldots n, \sum_{i=1}^{n} y_i \mu_i = 0
\end{aligned}
$$

$$
\begin{aligned}
w^* &= \sum_i \mu_i^* y_i x_i = \sum_{i \in S} \mu_i^* y_i x_i \\
S &= \{x_i \, \mu_i > 0\}, \text{ support vectors} \\
\mu_i^* &\left[1 - y_i(x_i^T w^* + b^*)\right] = 0 \forall i \\
\implies & 1 - y_i(x_i^T w^* + b^*) = 0 \\
\implies & y_i(x_i^T w^* + b^*) = 1
\end{aligned}
$$

Observations:

1. The hyperplanes that maximize the margin pass through some datapoints
2. These datapoints are called support vectors

8 SVM for not linearly separable case (14/03/2023)

$\nexists w$ s.t. $y_i(w^T x_i + b) > 1$

Introduce another variable into optimization $y_i(w^T x_i + b) > 1 - \xi_i$ (slack variable),

this allows misclassifications This can lead to all misclassification as slack variable leading to undesirable w Hence we also ξ_i to the optimization function

Primal:

$$\min_{w} \frac{1}{2}w^T w + \sum_{i=1}^{n} c\xi_i$$
$$\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i \text{ , } \xi_i \geq 0 \, \forall i$$

$$L(w, b, \xi, \mu, \lambda) = \frac{1}{2}w^T w + \sum_{i=1}^{n} c\xi_i + \sum_{i=1}^{n} \mu_i(1 - \xi_i - y_i(w^T x_i + b)) \sum_{i=1}^{n} \lambda_i \xi_i$$

K.K.T conditions:

1. $\nabla_w L = 0 \implies w^* = \sum_i \mu_i y_i x_i$
2. $\nabla_b L = 0, \implies \sum_i \mu_i^* y_i = 0$
3. $\nabla_\xi L = 0 \implies \mu_i^* + \lambda_i^* = c$, $\forall i$
4. $1 - \xi_i - y_i(w^T x_i + b) \leq 0, \xi_i \geq 0$, $\forall i$
5. $\mu_i \geq 0, \lambda_i \geq 0$
6. $\mu_i(1 - \xi_i - y_i(w^T x_i + b)) = 0, \lambda_i \xi_i = 0$, $\forall i$

$$\phi(\mu, \lambda) = \inf_{w,b,\xi} L(w, b, \xi, \mu, \lambda)$$

Here we have a term $\sum_i (c - \mu_i - \lambda_i)\xi_i$ which either becomes unbounded ($\xi \to \infty$ if $c - \mu_i - \lambda_i < 0$) or gives a trivial solution ($\xi \to 0$ if $c - \mu_i - \lambda_i > 0$) Hence we assume $c - \mu_i - \lambda_i = 0$ to get a good solution

Dual problem: (16/03/2023)

$$\max_{\mu} \sum_i \mu_i - \frac{1}{2} \sum_i \sum_j \mu_i \mu_j y_i y_j x_i^T x_j$$
$$\text{s.t. } \mu_i \geq 0 \text{ and } 0 \leq \mu_i \leq c$$

Can be solved using SMO (Sequential minimal optimization)
For $x_i \in S_i$, $\mu_i > 0$, $\epsilon = 0$, $y_i(w^T x_i + b) = 1$
Large $c \to$ less misclassifications (bias - variance trade)

9 Kernel SVM

Suppose $x \in \mathbb{R}^2$ (not linearly separable), with $x = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$

$$\begin{aligned}
&g(x) = ax_i + bx_2 + cx_1x_2 + dx_1^2 + ex_2^2 + f \\
&z = \phi(x), \phi : \mathbb{R}^2 \rightarrow \mathbb{R}^6 :\rightarrow \text{ feature transformation} \\
&\phi(x) = \begin{bmatrix} 1 & x_1 & x_2 & x_1x_2 & x_1^2 & x_2^2 \end{bmatrix} \\
&g(z) = w^T z
\end{aligned}$$

SVM with transformations $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$, $d' >> d$
$\mathcal{D} = \{(z_i, y_i)\}_{i=1}^n$
Dual:

$$\max_{\mu} \sum_i \mu_i - \frac{1}{2} \sum_i \sum_j \mu_i \mu_j y_i y_j \phi(x_i)^T \phi(x_j)$$
$$\text{s.t. } 0 \leq \mu_i \leq c$$

Suppose $\exists$ a function $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ s.t. $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$

$$\implies \max_{\mu} \sum_i \mu_i - \frac{1}{2} \sum_i \sum_j \mu_i \mu_j y_i y_j k(x_i, x_j)$$
$$\text{s.t. } 0 \leq \mu_i \leq c$$

10 Kernels in general

Mercer's Theorem: Let $\overline{k}$ is a $n \times n$ matrix with $\overline{k_{ij}} = k(x_i, x_j)$ If $\overline{k}_{n \times n}$ is a PSD $\forall \mathcal{D}$, then k is called a valid kernel ($\exists$ a space $\mathcal{H}$ and mapping $\phi : \mathbb{R}^d \rightarrow \mathcal{H}$, s.t. $k(.) = \phi(.)^T \phi(.)$)

$\overline{k}_{n \times n}$ PSD $\implies \sum_{i,j}^n \alpha_i \alpha_j k(x_i, x_j) \geq 0 \; \forall \alpha_i, \alpha_j \in \mathbb{R}$

Examples of valid kernel functions

1. Polynomial kernel: $k_P(x_1, x_2) = (1 + x_1^T x_2)^P$
2. Gaussian kernel: $k_G(x_1, x_2) = e^{\frac{-\|x_1 - x_2\|^2}{\sigma^2}}$ (generally good for SVM)
3. Sigmoid kernel: $k_S(x_1, x_2) = tanh(ax_1^T x_2 + b)$

11 SVM Summary

Given $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$, choose a $k(.)$ $\mu^* = \arg\max_\mu \sum_i \mu_i - \frac{1}{2}\mu_i\mu_j y_i y_j k(x_i, x_j)$, s.t. $0 \leq \mu_i \leq c$ (SMO)

Store x^* over S

$$\begin{aligned} h(x) &= \sum_{i \in S} \mu_i^* y_i k(x_i, x) + b^* \\ &= \sum_{i \in S} \mu_i^* y_i e^{\frac{-\|x_i - x\|^2}{\sigma^2}} + b^* \end{aligned}$$

It looks like **GMM** and **parzen window estimator** but only over a few datapoints

12 SVM as ERM

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2} w^T w + \sum_{i=1}^n c\xi_i \\ &\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i \text{ , } \xi_i \geq 0 \,\forall i \end{aligned}$$

Given an w and b, ξ_i has to satisfy, $\xi_i \geq \max(0, 1 - y_i(w^T x_i + b))$

$$\begin{aligned} \implies &\min_{w,b} \frac{1}{2} w^T w + \sum_{i=1}^n c \max(0, 1 - y_i(w^T x_i + b)) \\ &\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i \text{ , } \xi_i \geq 0 \,\forall i \end{aligned}$$

This can be written as $\Omega(w) + c\hat{R}(h)$

$$\begin{aligned} \hat{R}(h) &= \sum_{i=1}^n]max(0, 1 - y_i(w^T x_i + b)) \approx \mathop{\mathbb{E}}_{P_{xy}} l(h(x), y) \\ l(h(x), h) &= \max(0, 1 - y_i h(x)) \end{aligned}$$

13 Tasks

- ☐ Read SMO (Sequential minimal optimization) paper [for solving SVM]