Statistical Machine Learning

Lecture 5: Regression Trees and K Nearest Neighbors

2022-23

Regression Trees

- Assume the data set $\mathcal{D}\{(\boldsymbol{x}_n, t_n)\}_{n=1}^{N}$ is composed by deviates $\boldsymbol{X}$ and $\boldsymbol{T}$ such that $\boldsymbol{X} = \boldsymbol{x} \in \mathbb{R}^d$ and $T = t \in \{0, 1\}$.
- In the most naive approximation to the problem we may assume that

$$p_{X_\ell, T}(X_\ell = x, T = t) = \sum_{s=0,1} \alpha_{s,\ell}\delta_{t,s}\mathcal{N}(x|\mu_{s,\ell}, \sigma^2_{s,\ell}),$$

$$p_{X_\ell}(X_\ell = x) = \sum_{s=0,1} \alpha_{s,\ell}\mathcal{N}(x|\mu_{s,\ell}, \sigma^2_{s,\ell})$$

with $0 \leq \alpha_0, \alpha_1 \leq 1$.

Regression Trees

- The parameters of the model are obtained by 'suitable means' (not discussed in here).
- Those:

$$p_{T|X_\ell}(T = t|X_\ell = x) = \frac{\sum_{s=0,1} \alpha_{s,\ell}\delta_{t,s}\mathcal{N}(x|\mu_{s,\ell}, \sigma^2_{s,\ell})}{\sum_{s=0,1} \alpha_{s,\ell}\mathcal{N}(x|\mu_{s,\ell}, \sigma^2_{s,\ell})}.$$

.

Regression Trees

- The conditional entropy associated to feature ℓ becomes:

$$
\begin{aligned}
H_\ell &= \int_{-\infty}^{\infty} \mathrm{d}w\, p_{X_\ell}(w) \left\{ \sum_{t=0,1} p_{T|X_\ell}(t|w) \log_2 \frac{1}{\sum_{t=0,1} p_{T|X_\ell}(t|w)} \right\} \\
&= \sum_{t=0,1} \int_{-\infty}^{\infty} \mathrm{d}w\, p_{X_\ell,T}(w,t) \left[\log_2 p_{X_\ell}(w) - \log_2 p_{X_\ell,T}(w,t)\right)] \\
&= \int_{-\infty}^{\infty} \mathrm{d}w\, p_{X_\ell}(w) \log_2 p_{X_\ell}(w) - \\
&- \sum_{t=0,1} \int_{-\infty}^{\infty} \mathrm{d}w\, p_{X_\ell,T}(w,t) \log_2 p_{X_\ell,T}(w,t).
\end{aligned}
$$

Regression Trees

- The second term can be computed as:

$$
\begin{aligned}
\text{2nd term} &= \sum_{t=0,1} \int_{-\infty}^{\infty} \mathrm{d}w \, p_{X_\ell,T}(w,t) \log_2 p_{X_\ell,T}(w,t) \\
&= \int_{-\infty}^{\infty} \mathrm{d}w \, \alpha_{0,\ell} \mathcal{N}(w|\mu_{0,\ell},\sigma^2_{0,\ell}) \log_2 \alpha_{0,\ell} \mathcal{N}(w|\mu_{0,\ell},\sigma^2_{0,\ell}) + \\
&+ \int_{-\infty}^{\infty} \mathrm{d}w \, \alpha_{1,\ell} \mathcal{N}(w|\mu_{1,\ell},\sigma^2_{1,\ell}) \log_2 \alpha_{1,\ell} \mathcal{N}(w|\mu_{1,\ell},\sigma^2_{1,\ell})
\end{aligned}
$$

Regression Trees

-

$$
\begin{aligned}
\text{2nd term} &= \alpha_{0,\ell} \log_2 \frac{\alpha_{0,\ell}}{\sqrt{2\pi\sigma^2_{0,\ell}}} + \alpha_{1,\ell} \log_2 \frac{\alpha_{1,\ell}}{\sqrt{2\pi\sigma^2_{1,\ell}}} - \\
&- \frac{\alpha_{0,\ell}}{2\log 2} \int_{-\infty}^{\infty} \mathrm{d}w\, \mathcal{N}(w|\mu_{0,\ell}, \sigma^2_{0,\ell}) \left(\frac{w - \mu_{0,\ell}}{\sigma_{0,\ell}} \right)^2 - \\
&- \frac{\alpha_{1,\ell}}{2\log 2} \int_{-\infty}^{\infty} \mathrm{d}w\, \mathcal{N}(w|\mu_{1,\ell}, \sigma^2_{1,\ell}) \left(\frac{w - \mu_{1,\ell}}{\sigma_{1,\ell}} \right)^2 \\
&= \alpha_{0,\ell} \log_2 \frac{\alpha_{0,\ell}}{\sigma_{0,\ell}} + \alpha_{1,\ell} \log_2 \frac{\alpha_{1,\ell}}{\sigma_{1,\ell}} - \frac{1}{2} \log_2 2\pi - \frac{1}{2\log 2}
\end{aligned}
$$

Regression Trees

- The entropy becomes:

$$H_\ell = \int_{-\infty}^{\infty} \mathrm{d}w \, p_{X_\ell}(w) \log_2 p_{X_\ell}(w) +$$
$$+ \alpha_{0,\ell} \log_2 \frac{\sigma_{0,\ell}}{\alpha_{0,\ell}} + \alpha_{1,\ell} \log_2 \frac{\sigma_{1,\ell}}{\alpha_{1,\ell}} + \frac{1}{2} \log_2 2\pi + \frac{1}{2 \log 2}.$$

- Observe that if $\alpha_{0(1)} = 1$ then $H_\ell = 0$.
- Also observe that the first term is an integral that has to be solved applying numerical techniques.

Regression Trees

- By choosing the feature $\ell^\star = \min_{1\le\ell\le d}\{H_\ell\}$ that produces maximal information gain, we perform the partition of the set by using the criterion

$$x_{\ell^\star} < \alpha_{0,\ell^\star}\mu_{0,\ell^\star} + \alpha_{1,\ell^\star}\mu_{1,\ell^\star}.$$

- Subsequent partitions are obtained by applying the same techniques.
- The error of classification can be computed, once the last partition is performed, by counting the number of misclassified points.

Density estimation

- The probability that a new feature vector $\boldsymbol{x} \in \mathbb{R}^d$, drawn from an unknown density function $p_{\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x})$, will fall inside some region $\mathcal{R} \subset \mathbb{R}^d$ is, by definition:

$$\mathcal{P}_{\mathcal{R}} = \int_{\mathcal{R}} \mathrm{d}\boldsymbol{w}\, p_{\boldsymbol{X}}(\boldsymbol{w}).$$

- If we have N data points drawn independently from $p_{\boldsymbol{X}}(\boldsymbol{X} = \boldsymbol{x})$ then the probability that K of them will fall within the region $\mathcal{R}$ is given by the binomial law:

$$p_K = \frac{N!}{K!(N-K)!} \mathcal{P}_{\mathcal{R}}^K (1 - \mathcal{P}_{\mathcal{R}})^{N-K}.$$

Density estimation

- Observe that

$$\sum_{K=0}^{N} p_K = 1,$$

- and

$$\sum_{K=0}^{N} p_K K = N\mathcal{P}_{\mathcal{R}}$$

$$\sum_{K=0}^{N} p_K K^2 = N^2\mathcal{P}_{\mathcal{R}}^2 + N\mathcal{P}_{\mathcal{R}}(1 - \mathcal{P}_{\mathcal{R}}).$$

Density estimation

- By considering the variable $\xi \equiv K/N$ we have that

$$\mathbb{E}_\xi[\xi] = \sum_{K=0}^{N} \int \mathrm{d}\xi \, \delta\left(\xi - \frac{K}{N}\right) p_K \xi = \mathcal{P}_\mathcal{R}$$

$$\mathbb{E}_\xi[\xi^2] = \sum_{K=0}^{N} \int \mathrm{d}\xi \, \delta\left(\xi - \frac{K}{N}\right) p_K \xi^2 = \mathcal{P}_\mathcal{R}^2 + \frac{\mathcal{P}_\mathcal{R}(1-\mathcal{P}_\mathcal{R})}{N}$$

$$\mathbb{V}_\xi = \frac{\mathcal{P}_\mathcal{R}(1-\mathcal{P}_\mathcal{R})}{N},$$

thus ξ is a quantity with a vanishing variance when $N \to \infty$. Thus $\mathcal{P}_\mathcal{R} \approx K/N$.

Density estimation

- If the density function $p_{\boldsymbol{X}}(\boldsymbol{X}=\boldsymbol{x})$ is continuous and does not vary much inside the region $\mathcal{R}$,then we can approximate:

$$\frac{K}{N} \approx \int_{\mathcal{R}} \mathrm{d}\boldsymbol{w}\, p_{\boldsymbol{X}}(\boldsymbol{w}) \approx V p_{\boldsymbol{X}}(\boldsymbol{w}_0)$$
$$p_{\boldsymbol{X}}(\boldsymbol{w}_0) \approx \frac{K}{NV},$$

where $\boldsymbol{w}_0$ is the 'centre' of the region $\mathcal{R}$ and $V \equiv \int_{\mathcal{R}} \mathrm{d}\boldsymbol{w}$ is the volume of the region $\mathcal{R}$.

Density estimation

- We assume K is fixed. Starting at $\boldsymbol{w}_0$, the position of a fixture in the data set, we grow a sphere centered at $\boldsymbol{w}_0$ until we have precisely K data points inside it (not counting the one at $\boldsymbol{w}_0$).
- The final volume of the sphere with K neighbors $V_f(\boldsymbol{w}_0)$ is used to compute the density:

$$p_{\boldsymbol{X}}(\boldsymbol{w}_0) \approx \frac{K}{NV_f(\boldsymbol{w}_0)}. \tag{1}$$

KNN

- We can make use of this density estimator to construct class-posteriors through the Bayes' Theorem.
- Suppose we have a data set $\mathcal{D} = \{(\boldsymbol{x}_n, t_n)\}_{n=1}^{N}$, $\boldsymbol{x}_n \in \mathbb{R}^d$ and $t_n \in \{0, 1\}$.
- Suppose also that the total number of data points classified with a 0 (1) is N_0 (N_1).
- Clearly $N_0 + N_1 = N$. Those the prior class-probabilities are $P_0 = \frac{N_0}{N}$ and $P_1 = \frac{N_1}{N}$.
- We then draw a hypersphere around the point $\boldsymbol{x}$ which envelopes K points irrespective of their class.
- Suppose inside the volume $V(\boldsymbol{x})$ we have $K = K_0 + K_1$ points.

.

KNN

- We can use (1) to define the densities inside each class:

$$p_{\boldsymbol{X}|T}(\boldsymbol{X} = \boldsymbol{x}|T = t) = \frac{K_t}{N_t V}.$$

- By Bayes' we have that:

$$\begin{aligned} P_{T|\boldsymbol{X}}(T = t|\boldsymbol{X} = \boldsymbol{x}) &= \frac{p_{\boldsymbol{X}|T}(\boldsymbol{X} = \boldsymbol{x}|T = t)P_t}{p_{\boldsymbol{X}|T}(\boldsymbol{X} = \boldsymbol{x}|T = 0)P_0 + p_{\boldsymbol{X}|T}(\boldsymbol{X} = \boldsymbol{x}|T = 1)P_1} \\ &= \frac{\frac{K_t}{N_t V}\frac{N_t}{N}}{\frac{K_0}{N_0 V}\frac{N_0}{N} + \frac{K_1}{N_1 V}\frac{N_1}{N}} = \frac{K_t}{K}. \end{aligned}$$

- To minimize the probability of misclassifying a new vector $\boldsymbol{x}$, it should be assigned to the class t for which the ratio K_t/K is the largest.