CS 419M Introduction to Machine Learning **Spring 2021-22**

Lecture 21: Feature Selection and Auto Learning

Lecturer: Abir De *Scribe: Group 1*

21.1 Problem Definition for Feature Selection

Now in regression we say that our objective is to minimise loss i.e.

$$\min_{w,\ ||w||_0<=k} \sum_{i \in D} l(w^T x_i, y_i)$$

Here, $x \in^d$. Where d is the number of features.

Now we want to reduce the number of features because-

1. **Efficiency**: We want our model to work faster, want less calculations and thus reduce the number of parameters/ features.
2. **Privacy**: We do not want to include certain features in the training process. Suppose a person's hobby is rock climbing. Then if the google tracks the location of the person and is somehow tracked by the insurance company then the insurance company would have higher premium for the person which is undesirable.
3. **Resource Constraint**: Suppose I want to run the application on my smartphone. Now the space I can allocate in my phone is low but I still want faster and better results. Thus I would want to have less number of features.

21.2 Naive Method

1. For a given k, find all set of combinations of features possible.
2. Find MSE for all the set of features.
3. Select the features with minimum MSE.

Now the problem is that there $\binom{d}{k}$ combinations possible which are exponent in $f(d)$. Thus there would be too many combinations to choose from.

21.3 Successive Algorithm

1. Take $k' = 1$. Find the feature w_{01} which minimizes the MSE.
2. Take $k' = 2$. In addition to w_{01} find w_{02} which minimises the MSE given w_{01} is already selected.
3. Thus find a single feature in ech step which minimises the MSE given the features shown in previous step are selected.
4. Repeat till k features are selected.

Thus our goal is to-

$$\min_{w,S} \sum_{i \in D} l(w^T x_i, y_i) \quad \text{given,} \quad |S| <= k \ \text{ and, } \ S \in \{1, 2, ...d\}$$

Here, S is the set containing the indices of features. Let,

$$F(S) = \min_{w,S} \sum_{i \in D} l(w^T x_i, y_i) \quad \text{given,} \quad |S| <= k \ \text{ and, } \ S \in \{1, 2, ...d\}$$

Note: $F(S \ U \ e) <= F(S)$

Thus our algorithm is-

1. Start with S(0) = ϕ
2. For step t = 0 to k-1
   Find e(t) for given $S(t)$ such that, $F(S(t+1) \ U \ e(t+1))$ is minimised.
3. Report S(k) and thus find the optimal k features in the data space.

The solution F satisfies the following properties-

1. Monotonicity
2. $$\frac{F(S \ U \ e) - F(S)}{F(T \ U \ e) - F(T)} \geq \alpha$$
   whenever $S \leq T$. This is known as weak submodularity. If $\alpha = 1$, it is called submodular.

Perturbing features is therefore analogous to perturbing data.

Let F^* be an F we get by our approach. F_{opt} is something we get by solving the original objective directly.

$$F^* = O(F_{opt})$$

$$F^* \leq \frac{F_{opt}}{1 - e^{-\alpha}}$$

Here, $\alpha \leq 1$
This is because-

$$\begin{aligned} F(S) - F(S \ U \ e) &\geq 0 \\ F(S) - F(S \ U \ e) &\leq \alpha(F(T) - F(T \ U \ e)) \\ &\text{if } \alpha > 1 \text{ and } T = S, \\ F(S) - F(S \ U \ e) &< 0 \end{aligned}$$

which is a contradiction.

21.4 Active Learning

$$(x_1, x_2, ...x_n)_{n=10^4}$$

Now we want to label all these 10^4 samples. But it is costly to label all of them manually. Give $|s| = 10^3$ to label. Use this to label all remaining samples.

In regression, $\sum_{i \in D} l(y_i, w^T x_i) + \lambda ||w||$ But we can't minimize with respect to w as y_i are unknown.

$$\begin{aligned} &minF(x) \\ &X \subset Y = 10k \\ &|x| < 1000 \end{aligned}$$

Now, we are assuming, $y_i = w^T x_i + \varepsilon_i$

If given y_i and x_i both:

$$\begin{aligned} var(w) &= f(x) \\ E[||w - w^*||]^2 &\longrightarrow E[w] = w^* \end{aligned}$$

$$\begin{aligned} Note : w &= (\Sigma x_i x_i^T)\Sigma x_i(y_i) \\ &= (\Sigma x_i x_i^T)\Sigma x_i(w^T x_i + \varepsilon_i) \\ Note : \varepsilon_i &\sim \mathcal{N}(0, 1) \\ w &= (\Sigma x_i x_i^T)\Sigma x_i(x_i^T w^T + \varepsilon_i) \\ [w^* x_i &= x_i^T w^*] \\ w &= w^* + V^{-1} \sum_i x_i \epsilon_i \end{aligned}$$

Thus,

$$
\begin{aligned}
E[w] &= w^* \\
E[\|w - w^*\|^2] &= E[\|y^{-1} \sum_i x_i \varepsilon_i\|^2] \\
E[\|w - w^*\|^2] &= Trace(V^{-1}) \\
\text{Note: } E(\varepsilon_i \varepsilon_j) &= 0 \text{ (no correlation)} \\
E[\|w - w^*\|^2] &= E[(\sum_i x_i \varepsilon_i)^T (V^T V^{-1})(\sum_i x_i \varepsilon_i) \\
&= Trace[(\sum_i x_i \varepsilon_i)(V^{-1})^2 (\sum_i x_i \varepsilon_i)] \\
&= Trace[(V^{-1})^2 (\sum_i x_i \varepsilon_i)(\sum_j x_j \varepsilon_j)] \\
&= Trace[(V^{-1})^2 (\sum_i x_i x_i^T] \\
&= Trace(V^{-1})
\end{aligned}
$$

21.5 Group Details and Individual Contribution

- (19D070003) Adit Akarsh: Section 21.3
- (190100007) Aditya Vijay Jain: Section 21.1, 21.2
- (200110048) Ishita Tyagi: Section 21.4