UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

MOE: Mixture Of Experts

Analisi della letterature e delle hierarchical MOE

Modellazione Statistica - Data Science

Ivan Diliso - 761053

Indice

1 Ricerca iniziale delle informazioni **3**
1.1 Differenze con altri metodi 3

2 Dettagli tecnici **3**
2.1 Apprendimento 4
2.2 Funzioni di errore 4
2.3 Errore con gaussian mixture 5
2.4 Errore con MLP-Experts 5
2.5 Hierarchical Mixture of Experts 5
2.6 Model selection 6
2.7 Applicazioni pratiche 6

3 MOE e HMOE in R **6**
3.1 MEteorits: Mixtures-of-ExperTs modEling for cOmplex and non-noRmal dIsTributions . . . 6
3.2 mixtools: Tools for Analyzing Finite Mixture Models 6

1 Ricerca iniziale delle informazioni

MOE studiati nel campo delle reti neurali rispetto ad altri metodi ensemble come combining classifiers e ensemble of weak learners. Si utilizza un metodo dividi et impera per addestrare una serie di modelli parametrici e unirli per avere una soluzione. Se nei classici ensemble ogni learner è addestrato sullo stesso task in ME utilizza il divide et impera per dividere un task complesso in sottotask e ogni esperto è addestrato su task differenti, il modello di gating serve ad unire le soluzioni. A differernza dei modelli ensemble non c'è la necessità di rendere i learner individuali diversi in quanto ogni learner è addestrato per un task diverso. Il problema da risolvere è infatti trovare una divisione naturale dei dati. Una metodologia base è di targettare ogni esperto ad una diversa distribuzine specificata dalla funzione di gating, rispetto che apprendere la distribuzione originale dei dati.

SOFT PARTITIONING DEI DATI

- Partizionamento INPLICITO: Feature space viene diviso implicitamente in sottospazi tramite una funzione di errore, gli esperti gli esperti si specializzano in ogni subspace. Approccio competitivo MILE (Mixture of the implicitlò localised experts)
- Partizionamento ESPLICITO: Si utilizza un algoritmo di clustering per il partizionamento dei dati, questi vengono poi assegnati ad un esperto MELE (MIxture of expliticly localised experts)

1.1 Differenze con altri metodi

Altri metodi producono esperti unbiased con stime di errori non correlati. ME produce esperti biased con stime correalte negativamente.

Necessaria conoscenza pregressa di una divisione dei dati IL DATASET DEVE ESSERE DIVISIBILE.

Nella version convenzionale

Descrizione Tecnica di ensemble learning

Funzionamento di base Decompongo il task del modello predittivo in più sotto task, addestrare un esperto di quello specifico task su ogni task per poi sviluppare un gating model in grado di apprendere che esperto richiamare in base all'input e come combinare le predizioni. Posso suddividere il feature space di input in più feature space e addestrare un modello su ognuno di essi. Approccio divide et impera. I problemi possono essere sovrapponibili, non sovrapponibili e esperti su problemi simili collegati tra loro possono contribuire agli esempi che sono fuori dalla loro area di esperienza.

Questo approccio associa quindi un diverso peso ad ogni esperto, questa tecnica può essere vista come una forma di voting dei modelli ensemble, dove però la capacità di voto può cambiare al variare dell'input.

I pesi determinati dal gating network sono assegnati dinamicamente al variare dell'input, MOE quindi apprende che porizone del feature space è learned da ogni esperto dell'ensemble. I classificatore individuali sono addestrati per diventare esperti in una pozione del featrue space. La funzinoe di gating quindi selezinoe che classificatore, pesato con la sua expertise utilizzare per ogni istanza.

- POOLING: Utilizzo solo il classificatore con il peso più alto
- COMBINING: Utilizzo una somma pesata degli output di tutti i classificatori

2 Dettagli tecnici

Combinazione dei modelli, tipologia multi expert (diversi learner che lavorano in parallelo) con approccio locale (learner selection) si utilizza un modello di gating che guarda l'input e sceglie che modello è responsabile per generare l'output.

Sia $x \in \mathbb{R}^n$ vettore di input e si T il numero di esperti modello e $h_1, \ldots, h_T$ gli esperti del modello e y variabile target. Dati W_i parametri dell'i-esimo esperto, questo prova ad approssimare la distribuzione di y

$$h_i(y|x; W_i)$$

La funzione di gating produce un set di coefficenti the pesano il contributo degli esperti, sia v_i vettore dei pesi della funzione di gating relativa all'i-esimo esperto e α paramatro dell modello di gating, insieme dei pesi relativi ad ogni espeto, il set di coefficenti prodotti dal gating:

$$\pi_i(x; \alpha) : \sum_{i=1}^{T} \pi_i(x; \alpha) = 1$$

Sulla base di queste probabilità partizioniamo lo spazio di input, diverse partizioni appartengono a diversi esperti. L'output del modello sarà quindi:

$$H(y|x; \psi) = \sum_{i=1}^{T} \pi_i(x; \alpha) \cdot h_i(y|x; W_i)$$

Nella fase di training il valore $\pi_i(x; \alpha)$ indica la probabilità che l'istanza x appaia nel traning set dell'i-esimo esperto. Mentre nella fase di testing definisce il contributo che h_i da alla predizione finale.L'output della funzione di gating può essere espresso tramite una softmax

$$\pi_i(x; \alpha) = \frac{e^{v_i} x}{\sum_{l=1}^{k} e^{v_l} x}$$

usata sia per classificazione che per regressione.

2.1 Apprendimento

Gating network alloca dati di training a uno o più esperti e se l'output è incorretto il cambiamento dei pesi e localizzato su questo esperto. Locale in quasnto i pesi di un esperto sono disaccoppiati dai pesi di un altr oesperto.

Può avvenire tramite:

- GRADIENT DESCENT: Particolarmente utile con i mixuture of multi layer perceptron experts. Addestramento utilizzando questa funzione tende ad assegnare un dato di training ad ogni esperto
- EXPECTATION MAXIMIZATION: Metodi ME cercano di risolvere due task, dato un esperto trovare la funzione di gating ottimale e data la funzinoe di gating addestrare ogni esperto a massimizzare le performance sulla distribuzione assegnata dalla funzione di gating, questo rende naturale l'utilizzo di un algoritmo di expectation maximization

2.2 Funzioni di errore

$$E = \|y - \sum_j g_j O_j\|^2$$

I pesi di ogni esperto sono cosi aggiornati sulla base di un errore ensemble totale che sulla base dell'errore dello specifico esperto. Questo permette un alto livello di cooperazione e tende a sfruttare quasi tutti gli esperti del modello (nessun esperto non contribuisce al problema) In questa funzione di errore si assume che che l'output del sistema sia una combinazione lineare degli output degli esperti locali, con il gating che determina la proporzione. Strong coupling dei pesi.

$$E = \sum_j g_j \|y - O_j\|^2$$

Pesi aggiornati su errori singoli, non assicura la localizzazione degli esperti.

2.3 Errore con gaussian mixture

Una misura di errore che tiene conto di entrambi i fattori è basata sulla negative log probability di generare l'output vector desiderato, se si assume una mixture di modelli gaussiani con $\sum$ matrice di covarianza

$$E_{ME} = -log \sum_j g_j e^{-\frac{1}{2}(y-O_j)^T \sum^{-1} (y-O_j)}$$

L'apprendimento di ogni esperto avviene sull'errore individuale, ma l'aggiornamento dei pesi per ogni esperto è propozionale all suo rateo di errore sull'errore totale. Questo permette la localizzazione degli esperti nei sotto spazio delel feature corrispondente

2.4 Errore con MLP-Experts

Ogni esperto è un MLP con un hidden layer che produce un output O_j in funzione dell'inpnt con funzione di attivazione sigmoidale. Apprendimento con backpropagation massimizzando la log likelihood dei dati i parametri.

2.5 Hierarchical Mixture of Experts

Rimpiazzo ogni esperto con un sisteam completo MOE in modo ricorsivo. Si decide la profondità della ricorsione, il tipo di esperto e il tipo di modello di gating. Questo sviluppo ricorsivo crea una struttura ad albero. Può essere interpretato come un albero di decisione con i gating model che definiscono i nodi di decisione. Questa tipologia di albero viene definita "soft decision tree" in quanto i gating model ritornano una distribuzione di probabilità sugli esperti vengono quindi esplorate tutte le path dell'albero con differenti probabilità predendo poi una somma pesata a livello di foglie dove il prodotto è uguale al prodotto dei valori di gating di ogni path per arrivare allla foglia. In questa tipologia di apprendimento ogni nodo implementa una modello lineare (o regressione logistica) invece del valore costante di un albero CART. Nodi terminali chiamati esperti e nodi non terminali sono i nodi di gating.

L'idea è che ogni esperto da una opinione sulla predizinoe e queste sono combinate dal modello di gating. Seguendo il formalismo definito in precedenza e finito I numero di nodi connessi al gate al livello di radice e J_i numero di nodi connessi all'i-esimo nodo gate, π_i ouput del gata al livello di radice e $\pi_{j|i}$ output dell j-esimo gate connetto all'i-esimo gate

$$H(y|x;\psi) = \sum_{i=1}^{I} \pi_i(x;\alpha_{\pi_i}) \cdot \sum_{j=1}^{J_i} \pi_{j|i}(x;\alpha_{\pi_{j|i}}) h_{ij}(y|x;W_{ij})$$

Boundaries soft, dati anche in più boundaries. I dati sono dati in input agli esperti che producono dei vettori di output che procedono nell'albero verso alto, vengono moltiplicati tra loro e sommati seguendo i vari livelli dell'albero.
[

]

Vantaggi Le boundaries tra regioni di foglie sono non sono più "hard" ma c'è una transizione graduale da una all'altra, portatndo ad uno smoothing della risposta.

L'uso di soft split permette di catturare situazioni in cui la transizione da una risposta alta a bassa è graduale.

2.6 Model selection

Ottimizzazione di iperparamtrei, depth e connessioni dell'albero. Simile alla model selection applicata ad alberi, modificata la funzione di valutazione di un branch dell'albero.

- Modelli growing: aggiungo layer all'albero e determno la profondità e numero di esperti
- Pruning modelli: Riduzione dei requirement computazionali. Parametri costanti ma considero le path più probabili. Pruno i banch meno usati

2.7 Applicazioni pratiche

Dove è utilizzato? DA FARE

3 MOE e HMOE in R

MEclusternet, flexmix, mixreg, mixtools, flexCWM, meteorist

3.1 MEteorits: Mixtures-of-ExperTs modEling for cOmplex and non-noRmal dIsTributions

3.2 mixtools: Tools for Analyzing Finite Mixture Models

Riferimenti bibliografici

[1] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

[2] Alpaydin, E. (2020). Introduction to machine learning. MIT press.

[3] Zhou, Z. H. (2012). Ensemble methods: foundations and algorithms. CRC press.

[4] Zhang, C., & Ma, Y. (Eds.). (2012). Ensemble machine learning: methods and applications. Springer Science & Business Media.

[5] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer.

[6] Masoudnia, S., & Ebrahimpour, R. (2014). Mixture of experts: a literature survey. The Artificial Intelligence Review, 42(2), 275.

[7] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural computation, 3(1), 79-87.

[8] Yuksel, S. E., Wilson, J. N., & Gader, P. D. (2012). Twenty years of mixture of experts. IEEE transactions on neural networks and learning systems, 23(8), 1177-1193.

[9] Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. Neural computation, 6(2), 181-214.

[10] Gormley, I. C., & Frühwirth-Schnatter, S. (2019). Mixture of experts models. In Handbook of mixture analysis (pp. 271-307). Chapman and Hall/CRC.