自然言語処理 NLP と機械学習 ML・人工知能 AI

基礎知識確認試験

氏名（　　）

＜入門編をクリアしたエンジニア対象＞

【初級編】用語知識　＜20 分間 100 点満点＞

（1）下記の括弧に言葉を埋めなさい。

①形態素解析とは、（　　）と品詞付けを行う処理である。

②構文解析とは、形態素を(連)文節に統合する処理と、(連)文節どうしを結び付ける（　　）関係処理である。

③文脈解析とは、主に照応解析を行う処理形態で、代名詞や省略された（　　）から（　　）を特定する処理である。

④意味解析とは、文の意味概念を特定する処理形態で、関係子に（　　）を付与する処理である。

⑤コーパスとは、人手で解析学習記号が付与されたテキストを言うが、テキストだけのコーパスを特に（　　）と言っている。

⑥シソーラスとは、単語の（　　）概念と（　　）概念を体系化した辞書のことである。

⑦オントロジーとは、シソーラスに（　　）並びに役割関係を付加した単語概念体系辞書のことである。

⑧文とは、単文並びに単文を複数並べた（　　）と複数の単文が入れ子になっている（　　）のことである。

⑨機械学習には、教師付き学習である教師付き回帰と教師付き（　　）と教師なし学習がある。

⑩教師付き学習法で、ゼロ値が多い教師データを（　　）データという。

⑪一般にディープラーニングなどで使われている差分学習法を（　　）という。

⑫教師なし学習法で、ノイズなどを取り出す法を（　　）学習という。

⑬特徴ベクトルなどを抽出する為にコンパクトにすることを（　　）という。

⑭単純だが頑強な学習法を（　　）学習という。

⑮データ二値分類法で有名な方法を（　　）分類という。

⑯簡単な分類法を組み合わせた分類法を（　　）という。

⑰階層的分類をクラス又はクラス層分類といい、平坦な分類法を（　　）という。

⑱データが連続又は離散で並んでおり、予測などに使われる解析手法を（　　）という。

⑲状態などの動的計画やモンテカルロ法や TD 学習、Q 学習法を総称して（　　）という。

⑳多値分類法で二次元マップ上にベクトル要素で自動的に分類する方法を（　　）という。

【中級編】アルゴリズムと方法論＜60 分間 100 点満点＞

（1）下記の括弧に言葉や式を埋めなさい。

1）Q&A などの入力をd次元ベクトル $x \in \Re^d$ とし、出力凸関数も $f(x) \in \Re^d$ とする最適回答を求めるモデルを考える。

入力ベクトルから回答ベクトルへ探索をしていくモデルを $x_{n+1} = x_n + \Delta x$ とする。　…①

この差分ベクトル $\Delta x$ は、探索移動する「方向と量」を表している。ここで①式の出力凸関数処理を Taylor 展開で近似する。

$$f(x+\Delta x) \approx f(x) + \Delta x^T \cdot \nabla f(x) + \frac{1}{2}\Delta x^T \cdot \nabla^2 f(x) \cdot \Delta x \quad \cdots②$$

但し、$\nabla f(x)$ はナブラベクトル、$\nabla^2 f(x)$ はヘッセ行列である。ディープラーニングでは、ナブラベクトルは差分ベクトル、ヘッセ行列は差分ベクトルの積行列になる。ここで置き換えを行う。

$$g_n = \nabla f(x_n) \text{ 、 } H_n = \nabla^2 f(x_n)$$

ここで②式の展開式を書き直すと、

$$h(\Delta x) \approx \left( \qquad \right) \quad \cdots③$$

差分 $\Delta x$ の最小値を求める為、両辺を（　　）で微分すれば、

$$\frac{\partial h_n(\Delta x)}{\partial \Delta x} = g_n + H_n \Delta x \quad \cdots④$$

④式がゼロになるとき③式が最小値をとるので、そのときの $\Delta x$ を求めると、

$$\Delta x = \left( \qquad \right) \quad \cdots⑤$$

この $\Delta x$ が探索方向量であるので、ここでステップ係数 $\alpha$ を求める

$$\alpha^* = \arg\min_{\alpha} f(x_n - \alpha \Delta x) \quad \cdots⑥$$

探索方向量 $\Delta x$ とステップ係数 $\alpha$ が求まったので、①式にステップ係数を付加して更新する

$$x_{n+1} = x_n + \alpha \Delta x \quad \cdots⑦$$

上記法は Newton 法だが、ふたつの問題点がある。

ひとつは、$H_n^{-1}$ が正定値が条件で、（　　）では収束しない。

もうひとつは、行列なので次元の階乗になる為、メモリがオーバーフローする場合がある。

そこで準ニュートン法である L-BFGS 法では、正定値行列 $B_n = H_n^{-1}$ に近似する。すると⑤式が

$$\Delta x = -B_n g_n \quad \cdots⑧$$

もともと、$\frac{g_{n+1} - g_n}{x_{n+1} - x_n} \approx H_n$ だったので、$B_n$ もこの条件を満たすと、

$$B_n(x_n - x_{n-1}) = (g_n - g_{n-1}) \qquad \cdots⑨$$
$$B_n s_n = y_n$$
但し、$y_n = g_n - g_{n-1}$、$s_n = x_n - x_{n-1}$

もともと…の式は、近似式 $g_n \cdot g_n^T \approx H_n$ でも求めることができる。更新式は、

（　　）　　…⑩

但し、$\rho_n = y_n s_n^{T^{-1}}$

この⑩式は、⑦式と等価である。ディープラーニングも隠れ層である行列 $B_n$ の更新と等価である。すなわち、特徴ベクトルを求める為のウエイト行列は、探索方向量ベクトル $\Delta x$ のナブラベクトル $g_n = \nabla f(x_n)$ から $B_n = H_n^{-1}$ を求めていくことであるので、

深層学習/ディープラーニングとは、1980 年代の「準ニュートン法」そのもので、かなり古い最適解手法であったことが判る。

2）ベクトル処理などでよく使われる L1 ノルム $\|w\|$ を凸集合にして収束(最適解)を可能にする技術を考える。下記の $w_i, \eta_i$ は両凸である。

$$\|w\| = \sum_{i=1}^{d} |w_i| = \frac{1}{2} \sum_{i=1}^{d} \min_{\eta} \left( \frac{w_i^2}{\eta_i} + \eta_i \right) \qquad \text{但し、} \eta \in \Re^d, \eta_i \geq 0 \qquad \cdots⑪$$

これは相加平均と相乗平均を使っており、証明はヘシアン行列で行う。

$f(w,\eta) = w^2/\eta + \eta$ のヘシアン行列は、

$$\begin{bmatrix} \frac{\partial^2 f}{\partial w^2} & \frac{\partial^2 f}{\partial w \partial \eta} \\ \frac{\partial^2 f}{\partial w \partial \eta} & \frac{\partial^2 f}{\partial \eta^2} \end{bmatrix} = \frac{2}{\eta} \begin{bmatrix} 1 & -\frac{w}{\eta} \\ -\frac{w}{\eta} & \frac{w^2}{\eta^2} \end{bmatrix} = \frac{2}{\eta} \begin{bmatrix} 1 \\ -\frac{w}{\eta} \end{bmatrix} \begin{bmatrix} 1 & -\frac{w}{\eta} \end{bmatrix}$$

なので、$w \geq 0, \eta > 0$ のとき、（　　）であることから凸（劣モジュラ）であることが確認できる。

この $\eta \in \Re^d, \eta_i \geq 0$ は、$w$ のパラメータになっており、確率条件式で書くと $p(w|\eta)$ と書ける。

ちなみに⑪式右辺の関数項は、相加平均と相乗平均の不等式の変形である。

（　　）から $\frac{1}{2}(x^2 + y^2) \geq xy$ とし、$\frac{1}{2}\left(\frac{x^2 + y^2}{y}\right) \geq x$ から $\frac{1}{2}\left(\frac{x^2}{y} + y\right) \geq x$

としたのが①式である。だから、（　　）が付いていて不等式が等号に近くなることを凸の条件にしている。

ノルム空間は、余弦定理と内積空間と同等であることを理解すれば凸で収束可能であることが判る。

3）学習などに使われるハイパーボリックタンジェントとシグモイド関数、ロジスティック関数の相関関係を考察する。

シグモイド関数は、$\varsigma_a(x)=\dfrac{1}{1+e^{-ax}}=\dfrac{\tanh(ax/2)+1}{2}$ である。

従って、これをハイパーボリックタンジェントに変形すれば、

$\tanh x=\dfrac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$ になり、tanh と sigmoid は（ ）関係であることが判る。

そして、ロジスティック関数は、

$N=\dfrac{K}{1+e^{rK(t_0-t)}}$ であるので、上記シグモイド関数は、ロジスティック関数の特殊ケースで、

$r=a, K=1, t_0=0$ とすればシグモイド関数になる。

r は増加率、K は収容力(飽和)、$t_0$ は原点(変曲点)になる。

すなわち、ロジスティック関数は、ゼロから K までの S 字関数で、これをゼロから1までにしたのがシグモイドである。

そして、これを－1 から＋1 までにしたのが tanh である。用途に応じて、確率値にするか、符号化にするか、はたまた生物や経済などの成長率（飽和状態が有するモデル）にするか。このモデルには微分方程式が隠れている。

そして、重要な事は、この関数をディープラーニングの学習法（勾配法）として採用していること。

$\tanh x=\dfrac{e^{x}-e^{-x}}{e^{x}+e^{-x}}=f(x)$ とすると、

$\dfrac{\partial f}{\partial x}=\dfrac{1}{coshx}=\left(\dfrac{1}{coshx}\right)^2=\left(\dfrac{2}{e^{x}+e^{-x}}\right)^2=($ $)=1-\left(\dfrac{e^{x}-e^{-x}}{e^{x}+e^{-x}}\right)^2=1-tanh^2x$

$=1-\{f(x)\}^2$

そして、シグモイド関数を簡単に $f(x)=\dfrac{1}{1+e^{-x}}$ とすると、

$\dfrac{\partial f}{\partial x}=\dfrac{e^{-x}}{\left(1+e^{-x}\right)^2}=\dfrac{1+e^{-x}-1}{\left(1+e^{-x}\right)^2}=\dfrac{1+e^{-x}}{\left(1+e^{-x}\right)^2}-\dfrac{1}{\left(1+e^{-x}\right)^2}=($ $)=f(x)\{1-f(x)\}$

このようにハイパーボリックタンジェントもシグモイドも偏微分すれば簡単な計算式になる。誤差の収束モデルを凸関数のシグモイドにしているのは、変曲点での勾配が大きくなるので、その逆数をとることでスムーズにウエイトを調整することができる為であるが、偏微分計算が簡単になるので勾配係数で収束が容易になるのは大変便利です。