情報リテラシ 第一

2023年度1Q 5c/6c（IL1） 木曜日

担当: 地引

TA: 増井

テーマ 6 の講義内容

- データ サイエンスとは
- 観測されたデータから母集団の傾向を導く
  - 相関係数/最小二乗法
  - 不偏性/最尤推定
- 同、将来を予測する
  - 知識ベース(エキスパート システム)
  - 学習ベース(ニューラル ネットワーク)
- 数理解析の演習(Excel)

数理解析の演習(Excel)は、
時間の許す範囲で行ないます。

第 3 回小テスト（成績評価の対象になります）

- 課題の掲示先：

  » 2023 年度情報リテラシ第一 5c/6c ページ

  → “2.課題” → “4.第3回小テスト”

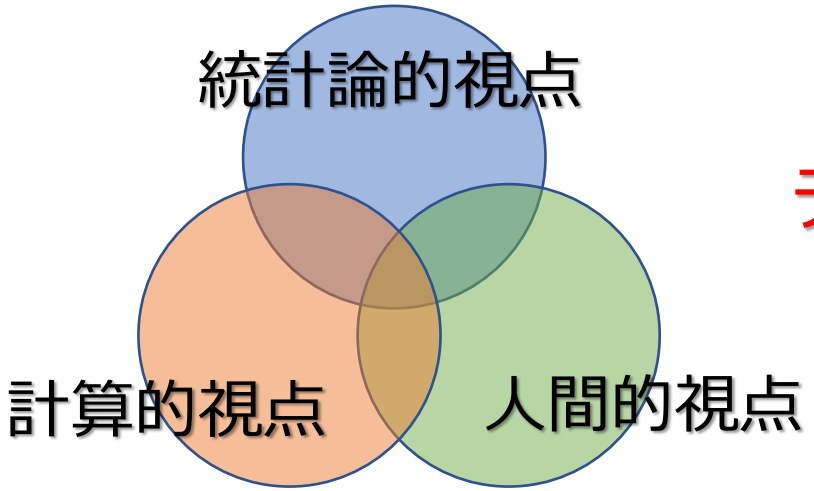
  → “情報リテラシ第一 課題「情報倫理とセキュリティ」” → Google Forms による回答

- **提出：6/7（水）まで（注意をよく確認して下さい）**

- 1Q の成績は、小テスト（3回）＋演習課題をもとに付けます。

データ サイエンスとは

- モノのインターネット（IoT： Internet of Things）
  - 世の中にある様々な物体(モノ)に通信機能を付与 + インターネットで接続
  - 様々なモノからデータを収集 → 収集したデータを利用したい（分析/制御）
- データ サイエンス
  - データに対して情報科学/統計学/アルゴリズムなどを横断的に用いて、新たな科学的/社会的知見を得る試み。



データを左図のような様々な観点から扱う

数理解析的視点

観測されたデータから母集団の傾向を導く

相関係数（1）

- 観測されたデータは、一般に (x, y) という形式
  - ある項目が x だった時、別の項目は y であった、という意味

- この時、x と y の関係を定量的に扱いたい。 最初の一歩
  - この定量性を、取り敢えず指標 C で表わすことにします。

- C に望まれる要件：
  - x が大きくなると y も大きくなる ⇒ C は正数 & その絶対値は関係の強さに対応
  - x が大きくなると y は小さくなる ⇒ C は負数 & その絶対値は関係の強さに対応
  - x の大小と y の大小に差がない ⇒ C は 0

相関係数（2）

- 値が大きい or 小さいの判断：平均値を基準
  - “$x - \underline{X}$”, “$y - \underline{Y}$” を考える（$\underline{X}, \underline{Y}$ は x, y の平均値）。

- x が大きい時は y も大きい ⇒ 指標 C は正数：積を適用
  - $(x - \underline{X})(y - \underline{Y})$ を考える。

- x が大きい時は y が小さい ⇒ C は負数：こちらも積は都合が良い

- x, y の関係の強さ ⇒ C の絶対値の大小：同上

相関係数（3）

- 指標 C の候補：$E[(x - \underline{X})(y - \underline{Y})]$

- (x, y) の傾向と他の観測データ群 (p, q) との傾向を比較したい。
  - 例えば、x は cm 単位、p は m 単位で計測した場合の扱いは?
    ≫データの単位を揃えるのは面倒。。
  - 性質や特徴が全く異なるデータ群同士の場合は?
    ≫片方が C = 30000 で、もう片方が 10000 となるより、0.3 と 0.1 の方が扱い易い。
    ≫例えば、0 ～ 1 の間に収まれば、確率的に扱える ⇒ モデル化に重要

- $-1 \leqq C \leqq 1$ となるように補正：下記が相関係数の計算式
  - $E[(x - \underline{X})(y - \underline{Y})] / \left(\sqrt{E[(x - \underline{X})^2]}\sqrt{E[(y - \underline{Y})^2]}\right)$

厳密な議論

観測されたデータとは

- 通常は、観測されたデータから平均や分散を計算している。
  - 分散：平均値から離れている量（これを⊿としましょう）の平均
    - ≫但し、単純に“⊿ = 平均値 − 観測値”としてしまうと、平均値からの大小に応じて⊿も±となり、⊿の平均を計算する際に相殺されてしまうので、実際の計算では、⊿の二乗の平均を計算している。
- ところでこの時、観測されたデータは対象全体（これを母集団と呼びましょう）の中で、どんな位置付けだろうか。
  - 母集団を全て観測できれば良いが、多くの場合は不可能
  - 例えば、クラス内の平均身長を得るために、クラス内の 10 人分だけ身長を測ったとして、「この選択された 10 人が偶然に背の高い人だった」という事象はないと言い切れるか?

観測値から求めた分散（1）

μ = 母集団の全データを対象とした平均（つまり、真の平均）

σ^2 = 同じく、全データを対象とした真の分散

$\hat{X}$ = 観測されたデータから計算した平均

以下、観測されたデータだけを用いて分散を計算してみると…

$$E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\hat{X}\right)^2\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}\left\{(X_i-\mu)-\left(\hat{X}-\mu\right)\right\}^2\right)$$

目がチカチカしますが、
真の平均 μ を基準に展開しています。

$$=E\left(\frac{1}{n}\sum_{i=1}^{n}\left\{(X_i-\mu)^2-2(X_i-\mu)\left(\hat{X}-\mu\right)+\left(\hat{X}-\mu\right)^2\right\}\right)$$

$$=E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right)-2E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\left(\hat{X}-\mu\right)\right)+E\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{X}-\mu\right)^2\right)$$

観測値から求めた分散(2)

計算が少々面倒ですが、前スライドの右辺3項を展開すると、
最後は下記のようになります。

$$E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\hat{X}\right)^2\right)=\sigma^2-2\frac{1}{n}\sigma^2+\frac{1}{n}\sigma^2=\frac{n-1}{n}\sigma^2$$

つまり、観測値だけから計算した分散は、真の分散になりません。

この問題は、どのように扱えばよいでしょうか。

不偏推定

- 母集団が持つ未知のパラメータ（値, 特徴）を S とします。
- また、n 個の観測データから計算したパラメータを S'_n とします。
- 観測データに偏りがない（不偏的である）場合、n を増やすにつれて S'_n も S に近付いて行くと考えられます。
- そこで、$E(S'_n) = S$ が言える時、この S'_n を真のパラメータ S と同じと考え、不偏推定量と呼ぶことにします。
- 問題は、どうやって不偏推定量を求めたらよいか、ということです。
  - まずは、観測データの誤差を手掛かりに考えて行きましょう。

確率密度関数（1）

結論から言うと、実は一筋縄では行きません。気合を入れましょう。

サイエンスなので、数理的に扱える構造（or 道具）が必要。

そこで、何か確率的な事柄が関係していると予想し、
ここから取り組んでみましょう。

まずは、次の確率 $P(Z)$ を考えてみます。

$P(Z)$ は、何となく $1/10$ になりそうな気もしますが、実は区間 $[1, 10]$ には数が無限に存在するので、例えば、$Z = 5$ になる確率 $P(Z = 5)$ は 0 になってしまいます。

「区間 $[1, 10]$ より一つの数値 Z を取り出す時、それが 5 になる確率は?」

よって、確率をもう少し意味があるように扱いたい場合は、
確率変数が連続値でも扱えるような定義が必要になる。

確率密度関数(2)

ある確率変数 Z が、a 以上 b 以下の値となる確率 $P(a \leqq Z \leqq b)$ を考えます。

事象によっては、$[a_1, b_1]$ と $[a_2, b_2]$ の長さは同じでも、
$[a_1, b_1]$ は選ばれ易く、$[a_2, b_2]$ は選ばれ難いという場合もある。

そこで、各 Z が選ばれる重み(or 割合)を $f(z)$で表わし、
確率 P を次のように定義し直します。

$$P(a \leq Z \leq b) = \int_a^b f(z)\, dz$$

この $f(z)$ を確率密度関数と呼び、以後は確率密度関数を中心に考えます。

改めて、観測されたデータとは

- 真の値に対して、誤差を含むデータ
  - ここで問題となるのは、どんな種類の誤差なのか？

- 誤差が本来的に持っていると想定される特徴
  1. 大きさの等しい正の誤差と負の誤差は、等しい確率で発生する。
  2. 小さい誤差は、大きい誤差より発生し易い。
  3. ある限界より大きな誤差は、ほぼ発生しない。

取り敢えずここでは、
「普遍的特徴」
と呼ぶことにします。

- 上記三つの特徴だけをもとに、一般的に誤差が従う確率（要は確率密度関数）を求めてみよう。

誤差が従う確率（1）

誤差が従う確率密度関数を f とする。以下では、この f 求めることを目指します。

まず、誤差が大きさが ε ～ $\varepsilon + d\varepsilon$ にある確率は下記で表わされます。

$$\int_{\varepsilon}^{\varepsilon+d\varepsilon} f(\varepsilon)\, d\varepsilon \xrightarrow{d\varepsilon \text{ は十分に小さいと考え、右で近似}} f(\varepsilon)\, d\varepsilon$$

母集団が備える真の値を X とし、観測データを x とすると、
$\varepsilon = x - X$ より上の確率は、下記のように書き換えることができます。

$$f(x - X)\, dx$$

n 回の観測結果が x_1 ～ x_n になる確率、即ち各誤差が ε_1～ε_n になる確率は、

$$f(\varepsilon_1)\, d\varepsilon \cdot f(\varepsilon_2)\, d\varepsilon \cdots f(\varepsilon_n)\, d\varepsilon = f(x_1 - X)\, dx \cdot f(x_2 - X)\, dx \cdots f(x_n - X)\, dx$$

誤差が従う確率（2）

ここで、n 回の観測結果が $(x_1, \ldots, x_n)$ となる可能性が最も高くなる場合を考えます。

これは、「真の値 X がどんな値の時、n 回の観測結果が $(x_1, \ldots, x_n)$ となる可能性が最も高くなるか」と言い換えることができます。

前に述べた誤差の本質的な特徴より、$n \to \infty$ を考えると、X が下記の場合の時、n 回の観測結果が $(x_1, \ldots, x_n)$ となる可能性が最も高いであろう。

（大きい値 x_i が観測される場合もあれば、小さい値 x_j が観測される場合もある。
しかし、誤差のバラツキが本質的特徴に従うならば、これらの平均値が真の値に最も近いだろう。）

$$X = \frac{x_1 + x_2 + \cdots + x_n}{n} \ (= \underline{x})$$

未知故に本当のところは所詮分からないので、このように仮定して、合理的な議論ができるか？

誤差が従う確率（3）

以上をまとめると、下記の関数 $P(X)$ は、$X = \underline{x}$ の時に最大値となる
（dx は一定値なので省略）。

$$P(X) = f(x_1 - X) \cdot f(x_2 - X) \cdots f(x_n - X)$$

$$P'(\underline{x}) = 0, \quad \underline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

もう少し詳しく言うと、
合成関数の微分を2回行ないます。

対数関数は単調増加なので、$\log P(X)$ は単調増加と思いがち
ですが、$P(X)$ が極値を持つならば $\log P(X)$ も極値を持ちます。

このままでは、$P(X)$ は掛け算が多くて扱い難いので、両辺の対数を取って微分します。

$$\frac{d \log P(X)}{dX} = -\sum_{i=1}^{n} \frac{f'(x_i - X)}{f(x_i - X)}$$

ここで $X = \underline{x}$ を代入すると、$\frac{d \log P(X)}{dX} = 0$ なので、

$f'/f = Q$, $x_i - \underline{x} = y_i$ と置き換えて、
右の関係式を得ます（$X = \underline{x}$ に注意）。

$$\sum_{i=1}^{n} \frac{f'(x_i - \underline{x})}{f(x_i - \underline{x})} = \sum_{i=1}^{n} Q(y_i) = 0 \quad (*1)$$

誤差が従う確率（4）

$\underline{x}$ の定義より、$\sum_{i=1}^{n} y_i = 0$ となるので(*2)、最終的には下記の問題に帰着します。

上の制約がある n 個の変数 y_i に対し、次の関係を満たす Q を求めよ。

$$\sum_{i=1}^{n} Q(y_i) = 0$$

正規分布は、「普遍的特徴」だけから導かれた分布です。
「普遍的特徴」は、様々な対象が備えていると予想されるので、
未知なるデータの分布を正規分布と仮定する場合は多いです。

どんな値 y_i を入れても微分値は 0 という意

以下、参考）

上式は、$Q(y_1) + Q(y_2) + \ldots + Q(y_{n-1}) + Q(y_n) = 0$ と表わされるが、*2より、$y_j\ (1 \leqq j \leqq n-1)$ が独立に動いて、y_n が y_j に束縛される（つまり、y_j に応じた関数値）と考えることができる（例えば 3 次元空間の x, y と z ）。

よって、両辺を y_1 で偏微分すると、$Q'(y_1) + Q'(y_n) \cdot \partial y_n / \partial y_1 = Q'(y_1) - Q'(y_n) = 0$ となる（*1より、$\partial y_n / \partial y_1 = -1$）。

以下同様に、$y_2, \ldots$ で偏微分して行くと $Q'(y_1) = Q'(y_2) = \ldots$ となるので、$Q'(y) = 0$ より $Q(y) = u \cdot y + v$ となります。

この式を*1に代入することで $v = 0$ が得られます。よって $f'(y)/f(y) = u \cdot y$ という形になるので、この微分方程式（変数分離形）を解くと、最終的に f は指数関数となります。 f は、一般的に**正規分布** or ガウス分布と呼ばれます。

最尤推定

- 誤差が従う確率を求める際の考え方（スライド 19 の赤字部分）
  1. 観測された n 個のデータが得られる条件付き確率を定義
  2. このデータの集合が、観測される可能性が最も高い組み合わせだと仮定
  3. 条件付確率を最大とする式/パラメータなどを計算（対数をうまく利用）
- これを最も尤もらしい推定という意味で最尤推定と呼びます。
- 最尤推定と不偏推定の関係
  - 最尤推定は考え方 or 計算方法などが分かり易いので広く普及しています。
  - しかしながら、必ずしも“最尤推定 = 不偏推定”とはなりません（反例あり）。
  - とは言え、どんな場合でも利用できる簡易な不偏推定法はありません。
  - よって、不偏推定の近似として最尤推定を利用しているのが現状です。

    ≫統計論的視点 → 真の値を求めると言うより、推定量の誤差を見積もる。

知識処理的視点

観測されたデータから将来を予測したい
人間ならば、どう考えるのか?

前期の知識処理

- 「IF 条件1 THEN 動作A ELSE 動作B」形式の知識がベース
  - もし 条件1 が成立するならば、動作A を実行せよ。さもなくば 動作B を実行せよ。
  - 例：IF 信号が青 THEN 進め ELSE 止まれ

- このような知識を数多く用意し、
  「現在の状況がどの条件に適合するかを調べて、対応する動作を実行する。」
  「これ↑を繰り返して行く。」

- コンピュータならば、条件の評価や動作を誤ることはないはずなので、
  正しく規則を作れば、絶対に間違いを起こさない知的判断ができる!!
  - このシステムは、エキスパート システム（専門家システム）と呼ばれました。

エキスパート システムの進化

- 知識を増やして行く必要あり。
  - 整理：知識には、事実（信号は赤/青がある）と規則（もし、赤ならば～）がある。
- 増えた規則から、適合する条件を高速に探し出せる必要あり。
  - 単に高速化するだけならば、並列検索で対応可能

但し、実際にエキスパート システムを作るとなると、以下をプログラムとして書きたい。

＊既にどんな知識が用意されているか?

＊その知識は、所望のものか?

＊上記がいずれも不可の場合は、新しい知識を適用

第一世代：真空管
第二世代：トランジスタ
第三世代：集積回路
第四世代：大規模集積回路

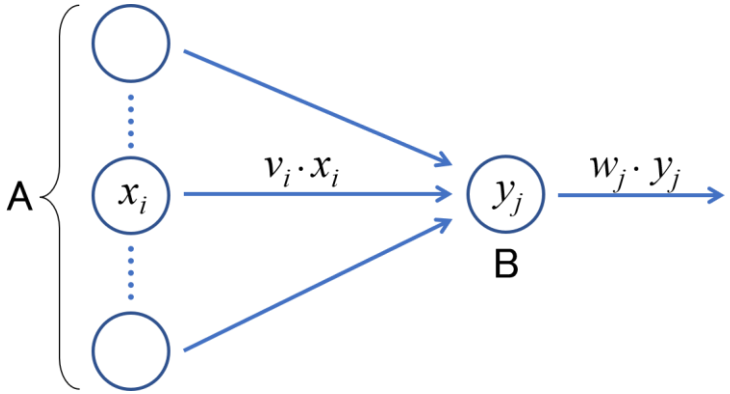
Prolog 言語（事実/規則/質問の3要素による推論 ＝ 総当たり探索）による第五世代コンピュータ

冬の時代、そして…

- エキスパート システムの行き詰まり
  - 当初は、推論により、新しい知識を自動的に蓄積して行けると考えていた。
  - 実際は、有益な知識は人間が入力（作成）する必要があった。
  - 単純なパターン マッチでは、有益な知識を得にくい。
    - ≫ 例えば、単純結合だけでは、意味のある結合と意味のない結合の区別がつかない。

どこかで見たな…
Web検索とか…

- そもそも、人間はどう考え、どう判断しているんだろう。
  - 人間の神経モデルを計算機上に作ってみたら、人間らしい判断はできるのか?



神経細胞群Aが持つ各電位 x_i に重み v_i が反映されて B へ伝わる。
B は受け取った v_1x_1 ～ v_nx_n を加算し、その結果 y_j へ重み w_j を反映して、隣接する神経細胞へ伝える。

これは、ニューロン/パーセプトロンと呼ばれています。

ニューラル ネットワーク（NN： Neural Network）

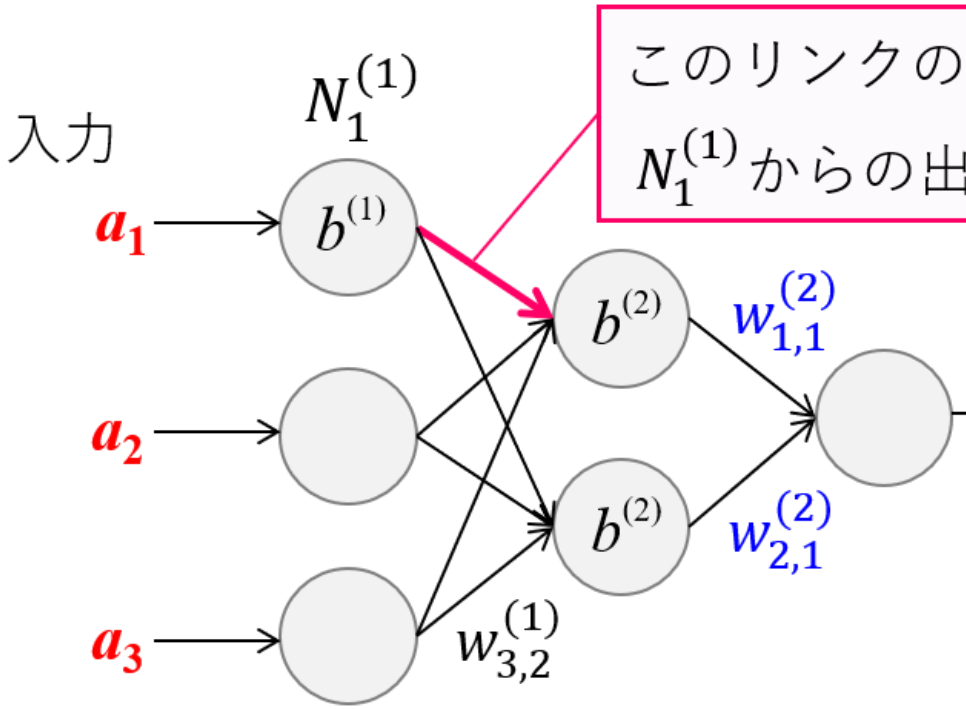
- 各ニューロンを様々な形態で接続

- 現実社会のデータを$(a_1, a_2, \ldots, a_n)$という数値ベクトルで表現
  - 社会には様々なデータがあるので、一つのデータ毎に数値ベクトルを作成

- 数値ベクトルを NN へ入力し、出力データを得る。
  - 神経モデルを作成したとして、人間は知識を蓄積して行くが…
  - 何かを契機として、重み v, w が変化して行くというのはどうだろう。

- 現実のデータから教師データを作成し、出力データと教師データを用いて、誤差逆伝播法によりニューロンを接続する各重みを修正して行く。

誤差逆伝播法（バック プロパゲーション）



このリンクの重み：$w_{1,1}^{(1)}$

$N_1^{(1)}$ からの出力：$u_{1,1}^{(1)}$

$b^{(i)}$：神経細胞 $N^{(i)}$ が持つ電位

$b^{(i)}$ も重み $w_{s,t}^{(i)}$ と同様に修正

$$u_{1,1}^{(3)} = y = \sigma(w_{1,1}^{(2)} \cdot u_{1,1}^{(2)} + w_{2,1}^{(2)} \cdot u_{2,1}^{(2)} + b^{(3)})$$

教師信号 t との誤差：$E = \frac{1}{2} \cdot (y - t)^2$

誤差の修正：$w_{j,k}^{(i)} = w_{j,k}^{(i)} - \alpha \frac{\partial E}{\partial w_{j,k}^{(i)}}$

E の中には y が入っており、y の中には各 $w^{(2)}$ が入っている。
これを各 $w^{(2)}$ で偏微分した値に応じて、その $w^{(2)}$ を修正して行く。

NN の特徴

- NN の数理的意味
  - 入出力は数値ベクトル + 電位と重みの線形結合 + 非線形関数 σ
    ⇒ NN は数学的関数（但し、“各重み = 係数” なので、超高次元関数）
  - 未知の事象（データ）であっても、過去の経験から大きく逸脱していなければ、この関数の何処かに乗っている or 近傍にあるはず。

- NN は、精度の良い予測が期待できる。
  - 但し、エキスパートシステムのような完全な演繹法ではない。
  - 過去の経験から得られた（or 誤差逆伝播法より得られた）近似関数

- とは言え、ニューロンの接続形態や非線形関数 σ 、数値ベクトルの作り方を工夫することで、飛躍的に予測精度が向上した。

知識ベースから学習ベースへ

- NN では、人間が知識を作っていない。
  - 学習（誤差逆伝播法）を繰り返すことにより、次第に重みが最適化
- AI の主流は、知識ベースから学習ベースに遷移して来た。
  - 但し、NN 内の重みを見ても、どんな判断が行なわれているのか不明
  - NN に懐疑的な側面
- NN = 数学的関数 ⇒ データを数値ベクトルで表現する必要あり
  - 将棋は棋譜を数値ベクトルとして表現し易い。
    ≫ 例えば、学習データ = (現在の局面, 次の一手), 教師データ = (最終的な勝敗) という感じ
  - では、文脈を表現する数値データとは?
  - 書割の人間か or 本当の人間か? ⇒ 自動運転の壁

今後の予定

- 6/15 (木) より、情報リテラシ第二が始まります。
  - 履修申告を終えていない人は、早急に申告して下さい。
    クラスについては、情報リテラシ第一と同じクラスを申告して下さい。
  - 1Q において情報通信に纏わる基本的な原理を理解したという位置付けで、
    2Q ではデータを解析してまとめる(表現する)演習が主体となります。

- 課題提出状況を確認して下さい。
  - 「提出締め切り」と表示されるまでに確認して下さい。

アンケートをお願いします(6/11 まで)。