深層学習

推論の信頼性

会津大学 コンピュータ理工学研究科 コンピュータ情報システム学専攻 髙橋輝

推論の不確かさ

入力$\mathbf{x}$から, 何らかの量yを予測する問題において, どのようド結果を信頼できるかを定量的に図ることができればうれしい.

そのためには, 確率分布$p(y|\mathbf{x})$を予測するようにすればよい.
$\mathrm{argmax_y p(y|\mathbf{x})}$を選べば一つの値を確定することができる.

多クラス分類では, 元々確率分布$p(C_k|\mathbf{x})$を予測するようになっている.
だが, 回帰では, $\mathbf{x}$からyの値一つを予測するようになっている.
この場合, $p(y|\mathbf{x})$をパラメトリック(いくつかのパラメータを定めることで再現できる)確率分布でモデル化することで, 推論の信頼性を考えやすくなる.

問題点

このようにして得られた分布の一つの予測でしかないため, その推論がどれだけ信頼できるものかは不透明.

不確かさの種類

不確かさを発生原因に基づいて分類

- 偶然性による不確かさ
- 知識の欠如による不確かさ

例

1. ある関数 $y = f(x)$ に従って生成された集合 $\mathcal{D} = \{(x_n, y_n)\}_{n=1,\cdots,N}$ が与えられ, $\mathcal{D}$から元の関数を推定したいとする.
2. 今, 真の関数が$y = \sin(x)$であるとして, 私たちが観測できるのは, それにノイズを加えた$y = \sin(x) + \epsilon_x$となる.
   (ここで, ϵ_xは, xに依存し, 正規分布$N(0, 0.005x^2)$に従うとする.)
3. 観測されるデータ$\mathcal{D}$は, 入力の範囲$0 < x < 2\pi$についてのみ得られる.

1-3の条件下で得られる点集合は, 図 8.1(a)のようになる.
このようなデータ$\mathcal{D}$が与えられたとき, yとx間の関係を表す適当なモデルを導入し,
これをデータにフィットさせる.
このとき, 図 8.2(b)のような予測が行われる(青色の領域は予測の精度)
このような予測の不確かさが, 偶然性による不確かさである.
次に, 知識の欠如による不確かさについて考える.
この種類の不確かさが起こる第一の要因として, 訓練データ$\mathcal{D}$が$0 < x < 2\pi$の範囲でしか与えられていないことがある(1).
この範囲の外側, $x < 0$もしくは$x > 2\pi$の領域ではデータが存在しないので, 不確かな予測しかできない.

もう一つの要因は, モデルが十分な表現能力を持っているかということだ. モデルの表現能力が足りなければ, 予測は正確にならない(2).
文献[1]では, (1)の不確かさを"近似由来の不確かさ", (2)を"モデル由来の不確かさ"と呼んでさらに区別している.

クラス分類の場合でも, 回帰分析の場合と同様に, "偶然性による不確かさ"と, "知識の欠如による不確かさ"の区別が可能.

- 偶然性による不確かさ
  2つのクラスが重なり合って存在しているとき, 自信をもって境界線を定めることができない.
- 知識の欠如による不確かさ
  サンプルがまばらにしか存在しないとき, 予測はやはり不確かになる. これはデータが十分でないことに由来する予測の不確かさである.

偶然性による不確かさは, どう頑張っても減らすことができないものととらえることができ, 条件設定が同じ限り, どう頑張っても減らすことができない.
知識の欠如による不確かさは, データの追加や, モデルの表現能力の改善によって減少させることができる.

とはいっても, 特徴量を追加することで, 偶然性による不確かさを減らすこともできる.
(特徴量エンジニアリングが大事な理由)

不確かさの数理モデル

ベイジアンニューラルネットワーク

知識の欠如による不確かさの確率分布による評価

知識の欠如による不確かさが$\mathcal{D}$の不足によるものならば, 与えられた$\mathcal{D}$からパラメータ$\mathbf{w}$をどれだけ曖昧さなく決定できるかに関係する.
$p(\mathbf{w}|\mathcal{D})$の分布の形が鋭い形を持つなら, $\mathbf{w}$は曖昧さなく決定できる.
$\rightarrow p(\mathbf{w}|\mathcal{D})$は知識の欠如による不確かさを表現した確率分布となっている.

ここで, 以前に$p(y|\mathbf{x})$を予測することで, 推論の不確かさを捉えるアプローチを紹介したが, この確率分布では, 一つの$\mathbf{w}$を一つ指定して, 入力に対して出力が決まるため, 知識の欠如による不確かさは含んでいない.
そこで, $\mathcal{D}$から決定できる$\mathbf{w}$のばらつきを考慮し, それぞれの$\mathbf{w}$での予測結果を平均(周辺化)すると, "$\mathcal{D}$を前提にあらゆる$\mathbf{w}$の可能性を織り込んだyの予測"を

$$p(y|\mathbf{x}, \mathcal{D}) = \int_{\mathbf{w}} p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|\mathcal{D}) d\mathbf{w}$$

のように表現できる.
これは, 2種類の不確かさを統合したものを表す.
ベイジアンニューラルネットワークは, $p(\mathbf{w}|\mathcal{D})$を使って, $\mathbf{w}$を周辺化することによって, 推論の不確かさを定量化しようとする試みだ.

最尤推定との関係

知識の欠如による不確かさは$p(\mathbf{w}|\mathcal{D})$で捉えられるが, 具体的に求めるのは簡単ではない.
今, "仮に", $p(\mathbf{w}|\mathcal{D})$が手元にあったとする.
このとき, 最低なパラメータ$\mathbf{w}$を求めるためには,

$$\hat{\mathbf{w}} = \mathrm{argmax}_{\mathbf{w}} \log \mathrm{p}(\mathrm{y}|\mathcal{D})$$

のように$p(\mathbf{w}|\mathcal{D})$を最大化すればよい.
ベイズの定理により,

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathbf{w}, \mathcal{D})}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

と書き直せるので,

上記の最大化は, 以下のように書き直せる.

$$\hat{\mathbf{w}} = \mathrm{argmin}_{\mathbf{w}}(-\log\ p(\mathcal{D}|\mathbf{w}) - \log\ p(\mathbf{w}) + \mathrm{const.})$$

ここで, $p(\mathbf{w})$は$\mathbf{w}$の事前分布である.
$-\log p(\mathcal{D}|\mathbf{w})$を$\mathbf{w}$について最小化することは, $\mathbf{w}$の最尤推定と一致する.
$-\log p(\mathbf{w})$は, 正則化を行うものと考えることができる.

ベイズ推定の立場から見れば, 重みの減衰項による正則化は, \textbf{w}の事前分布
$p(\mathbf{w})$に正規分布 $p(\mathbf{w}) \propto \exp(-\lambda/2||\mathbf{w}||^2) + \mathrm{const.}$を指定するのと同じ.

$$\log p(\mathbf{w}) = -\lambda/2||\mathbf{w}||^2\ + \mathrm{const.}$$

不確かさの予測

回帰の場合

分布$p(y|\mathbf{x}, \mathbf{w})$を予測するようにするには, ネットワークの損失関数を設計しなおす必要がある.
$p.3$で考慮したように, 予測すべき$p(y|\mathbf{x}, \mathbf{w})$を, 何らかのパラメトリックな確率分布, たとえば正規分布$y \sim N(\mu, \sigma^2)$でモデル化する. 正規分布は, μ, σで表現できるので, ネットワークはこれらを予測するように設計する.
(式は, 本に書いてあるものそのまま。口頭で解説します.)

学習は, これまで同様, 訓練データ$\{(\mathbf{x}_n, y_n)\}_{n=1,\dots,N}$の尤度を最大化することで行う.
なお, σはσ^2の形でしか尤度の式で現れないこと, またσ^2が非負であることを考慮して, $s(\mathbf{x};\mathbf{w}) \equiv \log \sigma^2$を新たに定義し, σのかわりに, これをネットワークの出力とする.

クラス分類の場合

ネットワークは元々各クラスkの事後確率$p\left(y=C_k\middle|\mathbf{x},\mathbf{w}\right)$を予測するように設計されており, そのままそれを偶然性による不確かさの予測に利用できる.
また、分類はy_kが最大となるクラスになされるが, その最大値$\max_k y_k$は, 分類したクラス数の事後確率の予測値なので, この推論の確信度として解釈できる.
しかし, 確信度は本来あるべき値より高くなる傾向があることが, 経験的に知られている.
確信度が, "その分類結果が実際に正解である確率"であると考えると, ある分類結果の確信度がその分類が正しい確率と一致する条件は,

$$p\left(\mathrm{argmax}_k y_k=\bar{k}\,\middle|\,\max_k y_k=q\right)=q$$

と表現できる.
ただし, $\bar{k}$は正しいクラスとする.

このような一致・不一致の度合いを測る尺度に, *期待校正誤差*(expected calibration error, ECE)を用いる.