汎化と正則化

筆者　佐藤 怜
所属　筑波大学 情報科学類
日付　2017.5.24

概要

本稿では, 機械学習の分野で汎化の概念と結び付けて議論されることの多い正則化に着目し, 何故同じような訓練誤差なら (あるいは多少訓練誤差を犠牲にしても) モデルの係数が小さい方が汎化誤差 (計測時にはテスト誤差) が小さくなるのかについて考察します.

1　オッカムの剃刀

「ある事柄を説明するためには, 必要以上に多くのことを仮定するべきではない」とする指針を, オッカムの剃刀と呼びます. 機械学習の分野では, 正則化や次元の呪いと関連付けて紹介されることが多いように感じます. 実際, モデルの解釈容易性, 計算量削減の観点からはこの指針は有用です.

ただ, この言説には数学的な裏付けがあるわけではありません. 従って, モデルの係数が小さい方が汎化誤差を抑えられる理由として扱うには物足りません.

2　非一様事前分布の導入による正則化項の導出

ベイズ的な観点からは, モデルのパラメータに非一様な事前分布を導入することで, 誤差関数に自然に正則化項を追加することができます.

データとラベルの組の全体を D, パラメータを w とします.

パラメータに関する事後確率はベイズの定理を用いて

$$P(w|D) = \frac{P(D|w)P(w)}{P(D)} \tag{1}$$

対数をとると

$$logP(w|D) = logP(D|w) + logP(w) - logP(D) \tag{2}$$

ここで, 求めたいパラメータ w は分布ではなく点の場合が多いので, 最大事後確率推定 (Maximum a Posteriori : MAP 推定) を行います. また, $-logP(D)$ の項は無視出来るので,

$$\arg\max_{w} \ logP(D|w) + logP(w) \tag{3}$$

ここで, $logP(D|w)$ はモデルの正しさなので, 誤差関数 Er で置き換えて

$$\arg\min_{w} \ Er(D, w) - logP(w) \tag{4}$$

$-logP(w)$ の最小化は, $logP(w)$ についての最大化と同じです. すなわちこの項は, 誤差関数に対して事前確率として選んだ確率密度関数 $P(w)$ が大きくなるような w(例えば $P(w) = N(0, \sigma^2)$ なら $w = 0$) を好むようにバイアスをかける働きをします.

ここで事前分布に平均 0 の正規分布を選んで最小二乗法で多項式回帰を解くと, 誤差関数には L2 正則化項が出現します.

しかし, この視点には既に平均が 0 の事前分布の方が, すなわちパラメータの絶対値は小さい方が良い結果が導ける筈であるという事前知識を導入しているため, 「何故小さな係数だと汎化性能が上がるのか」という疑問には答えていません.

準備: 汎化誤差最小化とマージン最大化

ここで, 汎化誤差の上界の最小化とマージン最大化が同値であることを紹介し, 後の章の準備を行います.

定理

データ数を n, データ分布を取り囲む超球の半径を r, ξ をスラック変数を纏めたベクトル, モデルのマージンを m とすると, ソフトマージン超平面識別モデルの汎化誤差 ϵ は次のような上界を持つ.

$$\epsilon \leq O(\frac{r^2 + ||\xi||^2}{nm^2}) \tag{5}$$

この定理から, マージンの最大化が汎化誤差の上界の最小化問題と同値であることがわかります.

3　マージン最大化とモデル勾配

この章では, マージン最大化が正則化に対応していることを示します.

3.1　サポートベクトルマシン

ハードマージンサポートベクトルマシンのラグランジュ関数 (未定乗数 λ) の最小化問題は以下で与えられます.

$$\min_{w,b} \sum_i \lambda_i (1 - y_i(wx_i + b)) + \frac{1}{2}||w||^2 \tag{6}$$

初項を詳しく見ると, まず $wx_i + b$ は w を法線ベクトル, b をバイアスとする分離超平面からデータ x_i への符号付き距離です. それに正解ラベル $y_i \in \{-1, 1\}$ をかけています. 従って正解した時には符号は正になりますから, $(1 - y_i(wx_i + b))$ の最小化は, 出来るだけ多くのサンプルに対して正解することを目指します.

ここで初項は, 分離超平面からの符号付き距離が大きい方が小さい値になりますから, 法線ベクトル w のノルムは大きい方が良いと思うかもしれません. しかし第二項では $||w||^2$ の最小化が行われます. 従って, この最小化問題は出来るだけ多くのサンプルに正解する w, b の中で w が小さいものを選ぶ問題になります. また, 第二項は L2 正則化と見なせます.

これがマージン最大化の定式化ですから, 単一の分離超平面を用いる分類器の汎化誤差の最小化はすなわち L2 正則化と同値です.

3.2 Lipschitz 連続性

ここまでの Vapnik の研究で, 既に汎化と正則化が数学的に結び付けられていることを説明しました.

この節では, 筆者が改めて Lipschitz 連続性を用いて正則化と汎化の関係を捉えることを試みます.

ここでは機械学習に於ける回帰 (regression) と分類 (classification) 問題を考えます.

データベクトルを $x \in \mathcal{R}^n$, 教師データを t で表します.t は回帰では $t \in \mathcal{R}$, 分類では $t \in \{1, 2, 3...k\}$ とします. k はクラス数です. 各データベクトル x には対応する t が存在し, その組を (x, t) とおきます.

今, $\mathcal{R}^n$ 上のベクトルに対応する t を生成する関数を考えます. すなわち手元にある学習データの組 (x, t) を生成した関数で, これは機械学習では神のみぞ知る真の関数です. 我々はこれを何らかのモデル $f(x)$ とデータ (x, t) で近似することを試みます.

さて, 関数が Lipshitz 連続であるとは, 関数が次の性質を満たすことをいいます.

$$\forall(x_1, x_2 \in \mathcal{R}^n) \quad \exists K > 0 \quad d_y(f(x_1), f(x_2)) \leq K * d_x(x_1, x_2) \tag{7}$$

これは, $f(x)$ の勾配が高々 K で抑えられることを意味します.

ここで, d_x という関数を用いました. これは距離の公理を満たす $\mathcal{R}^n$ 上の任意の関数です. また, d_y も同じく距離の公理を満たす任意の関数ですが, 回帰では $\mathcal{R}$ 上の関数とします. 分類では $\mathcal{N}$ 上の関数とし,

$$\begin{cases} 0 & (y_1 = y_2) \\ 1 & (y_1 \neq y_2) \end{cases}$$

で定義します.

さて, モデル関数 $f(x)$ の勾配が K で抑えられるとします. すなわちこの関数は Lipshitz 連続です. この時, 関数の出力が $d_y(y_1,y_2) = 1$ を満たすとします. この 1 という尺度は分類モデルではクラスが異なることと定義しました. すると, この関数の勾配は高々 K ですから, 出力が 1 違う為には, 入力が $\frac{1}{K}$ 離れている必要があります.

このことから, 分類問題においては異なるクラスに属する 2 つのデータは少なくとも $\frac{1}{K}$ 離れているということが言えます. ここにもマージンの考え方を導入すると, 2 クラスを分類する境界には少なくとも $\frac{1}{2K}$ のマージンが存在します.

さて, モデルの誤差関数に正則化項を導入すると, パラメータの絶対値が小さくなることが知られています. パラメータは一般に, 入力変数にかかる係数ですから, 係数が小さいということはすなわち勾配が小さいことを意味します. 同じような訓練誤差である 2 つのモデルの勾配が K と $2K$ であるとするならば, マージンは $\frac{1}{2K}$ と $\frac{1}{4K}$ ですから, 勾配が小さい方がマージンが大きくなります. すなわち正則化によって係数を小さくすると, マージンが大きくなり, 汎化誤差が抑えられることになります.

参考文献

[1] 赤穂昭太郎, サポートベクターマシン,
http://www.ism.ac.jp/~fukumizu/ISM_lecture_2006/svm-ism.pdf

[2] 中川裕志, サポートベクターマシン,
http://www.r.dl.itc.u-tokyo.ac.jp/~nakagawa/SML1/kernel1.pdf

[3] 数原良彦, 計算論的学習理論入門-PAC 学習とか VC 次元とか-,
https://www.slideshare.net/sleepy_yoshi/prm11-715
[4] 篠原歩 宮野, PAC 学習,
http://www.shino.ecei.tohoku.ac.jp/~ayumi/papers/IPSJ_pac.pdf
[5] 櫻井彰人, 情報意味論 (9)PAC 学習と VC 次元,
http://www.sakurai.comp.ae.keio.ac.jp/classes/infosem-class/2007/09vc.pdf
[6] John Shawe-Taylor, On the Generalization of Soft Margin Algorithms,
https://pdfs.semanticscholar.org/e543/7518f5c007b21d5b0e782640406831ffa7b4.pdf
[7] T.P.Runarsson S.Sigurdsson, Generalization theory,
https://notendur.hi.is/tpr/tutorials/svm/notes/chapter4.pdf
[8] Stephane Mallat, Understanding Deep Convolutional Networks,
https://arxiv.org/abs/1601.04920
[9] Christopher M. Bishop, パターン認識と機械学習 上