14 データの整形

機械学習アルゴリズムにデータを渡す際には、データをアルゴリズムに適した形に整形することが必要です。ここでは、代表的なデータ形式であるカテゴリ・数値データの整形について説明します。

カテゴリデータの整形

カテゴリデータは、性別や住んでいる地域など、そのデータのカテゴリを表しているものです。カテゴリデータは処理しやすいように数値に変換されますが、メモリ使用量や学習速度を考慮して様々な変換手法が提案されています。

なお、カテゴリデータを変換した後の数値の大小を比較することはできません。番号を割り振っただけで、その数の大きさに意味はないからです。

・ラベルエンコーディング

ラベルエンコーディングはもっとも単純な手法で、各カテゴリにひとつの数字を割り当てます。

・カウントエンコーディング

カウントエンコーディングは、そのカテゴリデータが出現した回数を割り当てます。

・One-Hotエンコーディング

One-Hotエンコーディングは列の名前をカテゴリ名にし、一致した列には1、それ以外の列には0を当てはめます。この場合、カテゴリの個数分だけ列の数が増えることになります。One-Hotエンコーディングはそれぞれのカテゴリを明確に分けることができます。しかし、列数が増えるためにメモリ使用量が増え、計算速度が遅くなってしまいます。

■カテゴリデータの整形

ID	都市
1	東京
2	大阪
3	名古屋
4	大阪

ラベル
エンコーディング

ID	都市
1	1
2	2
3	3
4	2

ID	都市
1	東京
2	大阪
3	名古屋
4	大阪

カウント
エンコーディング

ID	都市
1	1
2	2
3	1
4	2

ID	都市
1	東京
2	大阪
3	名古屋
4	大阪

One-Hot
エンコーディング

ID	東京	大阪	名古屋
1	1	0	0
2	0	1	0
3	0	0	1
4	0	1	0

次のページでは、数値データの整形方法を紹介します。データの値が数値であれば、通常整形をせずにアルゴリズムに渡すことができます。しかし、使用する機械学習アルゴリズムに適した変換を行うと、アルゴリズムがより高い性能を発揮することがあります。