

○ 訓練誤差

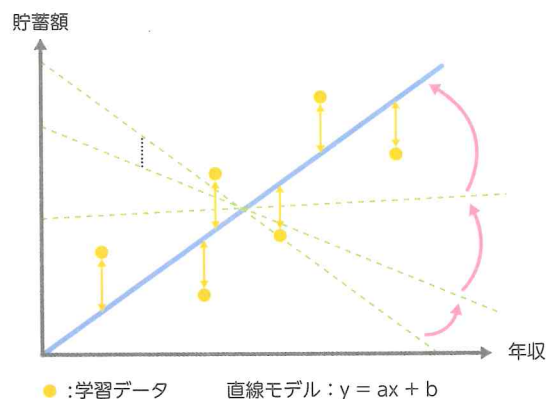
それでは実際に、入力データにフィットするような直線モデルのパラメータについて考えていきましょう。このようなパラメータを探すには、どのような処理が適切なのでしょうか。

まず、モデルに適当なパラメータを入れて出力を計算します。すると、必ず出力データと正答データの間に大きな差が見られます。この差のことを、**訓練誤差**といいます。学習においては、この訓練誤差が小さくなるようにどんどんパラメータを更新していくのです。たとえば傾きに関しては、下図のようにパラメータを更新することにデータに沿うように直線の傾きが変化していくことになります。

モデルとデータとの差が十分に小さくなったら、学習を終了します。こうして決めたモデルは、たとえば「年収が650万円」という今まで学習してこなかった未知の入力データに対しても、「貯蓄額は700万円」と予測できるのです。

今回は直線モデルを例として説明しましたが、機械学習で利用されるモデルにはもっと複雑なものも数多くあります。また入力データも年収だけではなく、年齢や世帯人数など他のデータも見て予測のモデルを作成するケースもあります。とはいえ多くの場合、教師あり学習における学習とは、このように「モデルの出力と正答データとの差が小さくなるようにパラメータを更新していくこと」と理解しておけばよいでしょう。

■ 直線モデルにおける学習



○ くり返し、学習する (イテレーション)

すべての学習データを何度もくり返し学習させることで、モデルのパラメータが調整されていき、徐々に正しい予測・分類結果を出力できるようになります。このくり返しのことを**イテレーション**といいます。イテレーションの方法として、主に「**バッチ学習**」「**ミニバッチ学習**」「**オンライン学習**」が存在します。

バッチ学習では、1回の学習ですべての学習データを読み込みます。それに対してミニバッチ学習では、1回の学習で「バッチサイズ」として設定した数のデータを、オンライン学習では1回の学習で1つのデータのみを読み込んで、学習を実施します。

バッチ学習ではすべての学習データを1度に処理する必要があるため、必要となるコンピュータのメモリの大きさが増加しますが、すべてのデータを均等に扱えます。対してミニバッチ学習やオンライン学習では、すべての学習データの中からランダムに、一部または1つのデータを読み込んで学習し、これをくり返します。結果的にすべてのデータを読み込むことにはなりますが、学習結果はより最後のほうに読み込んだデータに引っ張られるため、学習の順序によって性能が変わるケースがあります。また、バッチ学習に比べて学習を行う回数が多くなるため、計算量が大きくなる可能性が高いと言えます。バッチ学習とオンライン学習については、次のSectionでさらに詳しく解説します。

■ くり返し学習 (イテレーション) とその種類

バッチ学習



ミニバッチ学習



オンライン学習



まとめ

■ モデルの学習では、学習をくり返し訓練誤差を小さくする