機械学習　レポート

機械学習　要点まとめ

・線形回帰モデル

　最小二乗法を用いて線形近似することは、エクセル等でもなじみがある。機械学習のアルゴリズムとして重要なのは、二次元空間から拡張しても用いることができることで、行列を用いた偏微分方程式を解くことで、誤差が極小を取る重みを決定することができる。

・非線形回帰モデル

　曲線とフィッティングすることができる一方、正しくモデルを設定する必要があるため、扱う側にも十分な知識が必要である。また、表現力の高いモデルを用いるほど過学習の危険があるため、正則化項を用いたり、パッチで学習したりする際は、学習誤差と訓練誤差の推移を確認したりすることが重要である。また、ある程度モデルが決まれば、交差検証等も有効である。

・ロジスティック回帰モデル

　回帰問題と異なるのは、0か1に分類するところだ。線形回帰問題の操作に加えて、シグモイド関数によって確率として評価し、その確率が最も訓練データと近い重みを導出することでモデルを作る。これは、あるデータの分布をベルヌーイ分布等のモデルで仮定することにより尤度関数を求め、その関数の最大を取る重みを取ることで計算され、対数を取り正負を逆転することにより、最小二乗法を用いることができる。データの数が多い場合、全てのデータで重みを更新せず一部のデータだけで更新することを繰り返すことで、徐々に最小に近づける方法もある。

　性能の評価には様々な指標があり、取り方によって数値が大きく変わる場合もあるため、何を評価しているのかをしっかりと確認する必要がある。

・主成分分析

　特徴量を空間上にマッピングしたときに、広がりが大きい方向（分散が大きいベクトル）を探し出し、射影することによって、データがもっている特徴を要約する。学習する特徴量が多い時に、次元を圧縮することができる。

・アルゴリズム

　k 近傍法はあらかじめラベルがわかっているデータを用意して、分類したいデータに対して空間的に近い数点のデータの多数決によって分類する手法である。

　k-meansは、教師がなくても分類でき、他のでーたとの空間的な距離が近いもので固まりを作ってラベリングする。モデルの評価が難しいが、一般的には同じくラベリングされたものどうしの密集度と他にラベリングされたものとの乖離度をあわせた指標を用いて評価される場合が多いようだ。

・サポートベクターマシン

　サポートベクターを探し出すために、分離平面からの距離が最小であるデータを探索する。一方で、そのデータに対して、その距離が最大となる分離平面を探さなくてはいけない。これを解くためにラグランジュ未定乗数法を用いる。サポートベクターを探し出すまでに勾配降下法で係数を決定するので、学習率や工ポック数を適切に設定する必要があった。現実には誤差無く完全に線形で分離できる問題は少なく、ある程度の誤差を許容するソフトマージンだったり、カーネルトリックを用いて線形分離できるように工夫する場合もある。