PRMLノート 第1章 序論

学習とは

問題設定

N個の観測値を並べた$\mathbf{x} = (x_1, \ldots, x_N)$と，それらに対応する観測値$\mathbf{t} = (t_1, \ldots, t_N)$を得たとき，新しい入力$\hat{x}$に対して精度良く出力$\hat{t}$を予測する．

多項式フィッティング

例えば多項式$y(x, \mathbf{w})$を用いてデータへのフィッティングを行う．パラメータ$\mathbf{w}$を最適化する仮定を**学習**と呼ぶ．

例えば**最小二乗法**で誤差を最小化する．

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

純粋な最小二乗法だけでは**過学習**の危険性がある．

確率論

基本法則

確率の加法定理

$$p(X) = \sum_{Y} p(X, Y)$$

確率の乗法定理

$$p(X, Y) = p(Y|X)p(X)$$

ベイズの定理

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$$

ここで，$p(Y)$は**事前確率**，$p(Y|X)$は**事後確率**.

期待値と分散

ある関数$f(x)$の，確率分布$p(x)$の下での平均値を**期待値**と呼ぶ.
xが離散変数なら

$$\mathbb{E}[f] = \sum_x p(x)f(x)$$

連続変数なら

$$\mathbb{E}[f] = \int_x p(x)f(x)dx$$

$f(x)$が平均値からどれくらいバラついているかを表すのが**分散**.

$$\mathrm{var}[f] = \mathbb{E}\left[f - E[f]\right] = \mathbb{E}[f^2] - (\mathbb{E}[f])^2$$

共分散は

$$\mathrm{cov}[x, y] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

ガウス分布

平均μ，分散σ^2を持つガウス分布は以下で定義される.

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

ガウス分布は規格化されていて，そのまま確率分布のモデルとして使える.

最尤推定による多項式フィッティング

平均が多項式$y(x, \mathbf{w})$で表せるガウス分布に従ってtが分布していると，**尤度関数**は

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | y(x_n, \mathbf{w}), \beta^{-1})$$

これでは小さすぎるので，通常は対数を取って和の形にして最大化する.

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

$y(x, \mathbf{w})$が関係している項は，最小二乗法で出てきたのと同じ.

情報理論

確率変数xの**エントロピー**は,

$$H[p] = -\sum_{i} p(x_i) \ln p(x_i)$$

で定義される．鋭いピークを持つ分布ではエントロピーが低く，なだらかなピークを持つときはエントロピーが高い（分子の分布をイメージ）.

連続変数のエントロピーは

$$H[p] = -\int p(x) \ln p(x) dx$$

で定義され，微分エントロピーと呼ぶ.
最大のエントロピーを持つ分布はガウス分布.