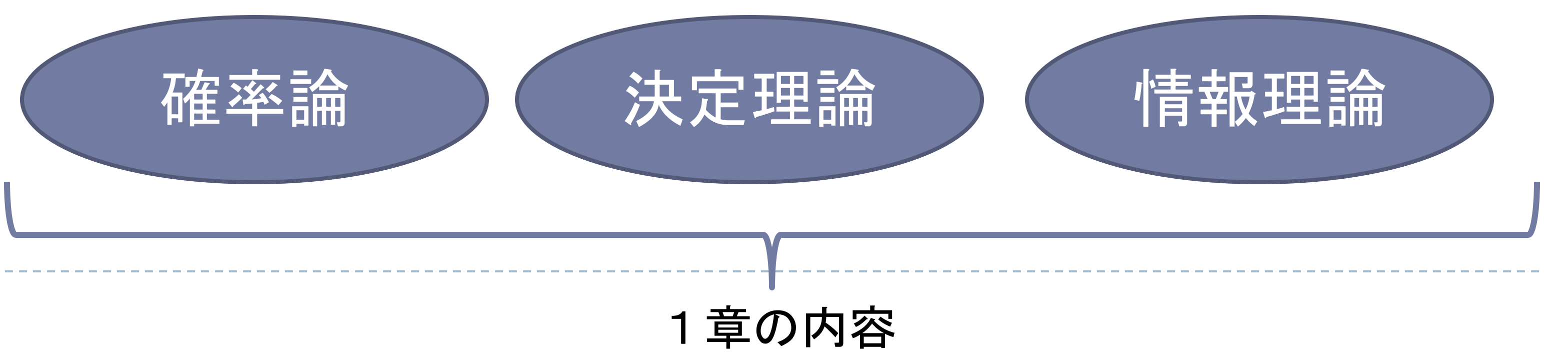
パターン認識と機械学習
第1章：序論（後半）

Christopher M. Bishop (2006):
Pattern Recognition and Machine Learning, Springer, pp.37-57

今回の内容と目次

- 1.5 決定理論
  - 1.5.1 誤識別率の最小化
  - 1.5.2 期待損失の最小化
  - 1.5.3 棄却オプション
  - 1.5.4 推論と決定
  - 1.5.5 回帰のための損失関数
- 1.6 情報理論
  - 1.6.1 相対エントロピーと相互情報量



決定理論

- 確率論：不確実性を定量化したり，操作を行ったりするための一貫した数学的枠組み
- 決定理論：パターン認識で遭遇する不確かさを含む状況に置ける最適な意思決定を行うこと
- 入力ベクトルxと対応する目標変数t
- 新たなxの値に対するtを予測することが目的
- ex)回帰問題：tは連続変数
- 　　クラス分類：tはクラスラベル
- 　　離散選択問題：tは離散変数
- 同時確率分布p(x,t)は変数に関する不確実性を完全に要約するもの
- p(x,t)を訓練データ集合から決めることが推論

決定理論の具体例（１）

- 具体例：医療診断問題：患者のX線画像で癌かどうかを判断する

- 入力ベクトルxは画像のピクセル強度
- 出力変数tは癌であるクラスC1(t=0), 癌でないクラスC2(t=1)

- 一般的な推論問題は同時分布p(x, t)を決定することであり,
- それが状況の最も完全な確率的記述である

- →これは非常に有用で情報量が多い記述であるが,
- 最終的には患者に治療を施すかどうかを決めなければならないので
- ある適当な基準の上で最適な決定をしたい
- これが決定(decision)の段階であり，適切に確率が与えられたときに
- 最適な決定をするにはどうするかを教えてくれる決定理論の主題である

決定理論の具体例（２）

- 新たな患者のX線画像xが得られたときに，画像を２つのクラスのどちらかに割り振ることが目標

$$p(C_k \mid x) = \frac{p(x \mid C_k)\,p(C_k)}{p(x)} \qquad (1.77)$$

$p(C_k)$ ：人が癌である/ない確率(これまでの観測(訓練集合)から得られる)

$p(x)$ ：xというX線画像である確率(これまでの観測から得られる)

$p(x \mid C_k)$ ：Ckのときにxというx線画像である確率(これまでの観測から～)

誤識別率の最小化（１）

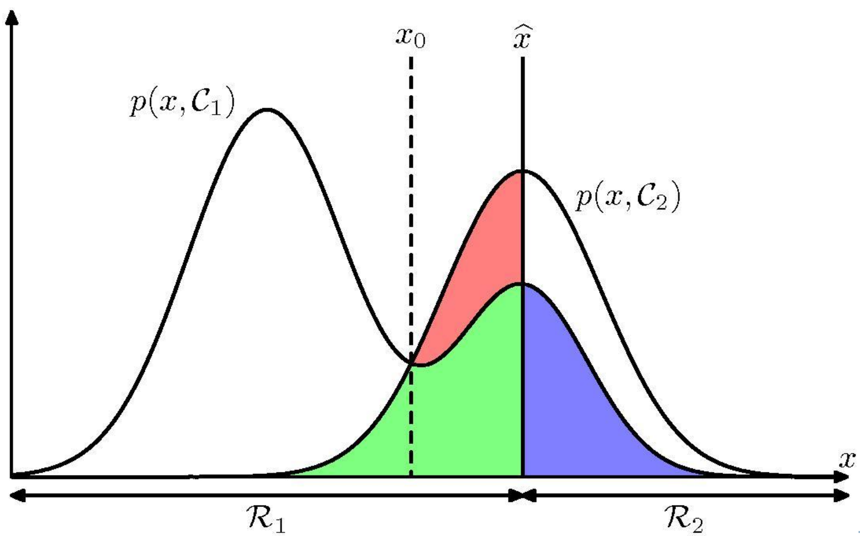
- (1)誤識別をできるだけ少なくすることを目標とする
- 決定のためにはxの各値に利用可能なクラスの１つを割り振るための規則が必要

- そのような規則は入力空間を各クラスに1つずつ対応する決定領域(decision region)と呼ばれる領域Rkに分割しRk上の点にはすべてクラスCkを割り当てる.
- 決定領域の間の境界は決定境界(クラス境界：decision boundary)あるいは決定表面(decision surface)と呼ばれる
- 各決定領域は連続とは限らず，いくつかの領域に分かれていることもあり得る.

誤識別率の最小化（２）

- 癌の例
- 誤識別とはクラスC1に属する入力ベクトルをC2に割り当てたりその逆が起きることである．それが起きる確率は

$$p(\text{誤り}) = p(x \in R_1, C_2) + p(x \in R_2, C_1)$$
$$= \int_{R_1} p(x, C_2)dx + \int_{R_2} p(x, C_1)dx \qquad (1.78)$$



誤識別を最小化するxの値はx0である

期待損失の最小化

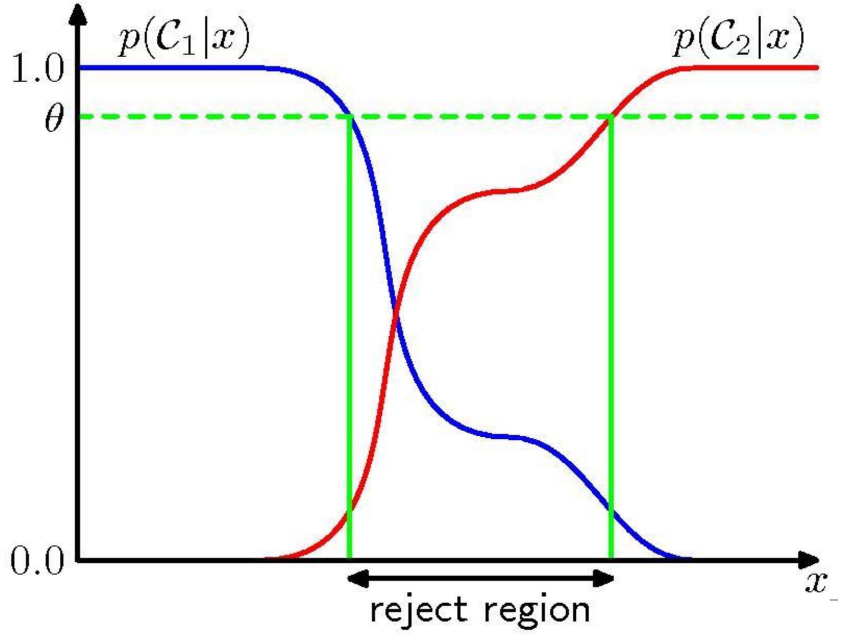
- （２）単純に誤識別を最小化すればいいのではない
- 正常な患者を癌と診断することと，癌の患者を正常と診断することの間には大きな違いが存在する
- そこで損失関数(loss function)，コスト関数(cost function)を導入

$$\mathbf{E}[L] = \sum_k \sum_j \int_{R_j} L_{kj}\, p(x, C_k)\, dx \qquad (1.80)$$

$$L_{kj} = \begin{array}{c|cc} & \text{癌} & \text{正常} \\ \hline \text{癌} & 0 & 1000 \\ \text{正常} & 1 & 0 \end{array}$$

棄却オプション

- すべてクラス分けするのが良いとも限らない
- 正確に分類できるところだけ自動的に分類し，曖昧なところは人（専門家）が行う方が全体のクオリティが向上する場合がある
- そのしきい値θを導入する



回帰のための損失関数（１）

- 曲線フィッティングの回帰問題の場合に戻る
- 決定段階は各入力xにおける特定の推定値y(x)を選ぶこと．その際，損失L(t, y(x))を被るとする．
- 平均損失は

$$\mathbf{E}[L] = \iint L(t, y(x))\, p(x,t)\, dx dt \qquad (1.86)$$

- 回帰問題の場合に良く使われる損失関数は二乗誤差であるので，

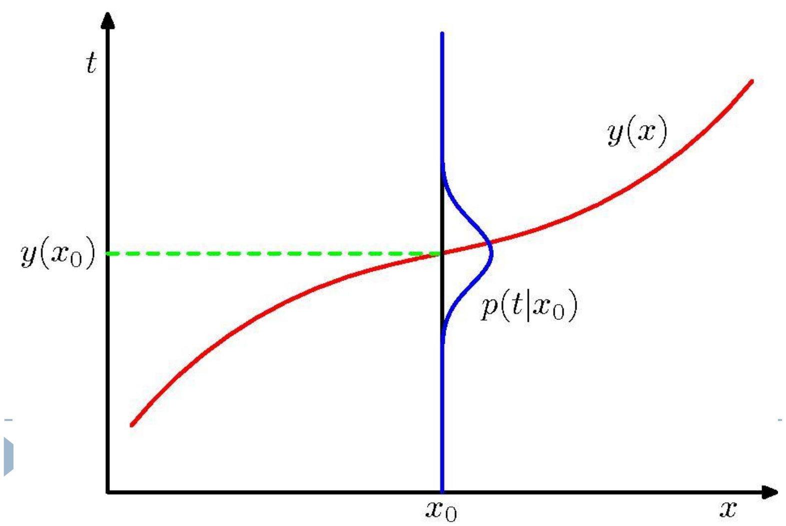
$$\mathbf{E}[L] = \iint \{y(x) - t\}^2\, p(x,t)\, dx dt \qquad (1.87)$$

回帰のための損失関数（２）

- 目標はE[L]を最小にするy(x)を選ぶこと
- 変分法を用いると

$$\frac{\delta \mathbf{E}[L]}{\delta y(x)} = 2\int \{y(x) - t\} p(x,t)\,dt = 0 \qquad (1.88)$$

$$y(x) = \frac{\int t p(x,t)\,dt}{p(x)} = \int t p(t \mid x)\,dt = \mathbf{E}_{\mathbf{t}}[t \mid x] \qquad (1.89)$$



これがよく知られた
回帰関数(regression function)

損失関数に二乗誤差を用いず
ミンコフスキー損失を用いた拡張版
もある

情報理論

- 離散変数xを考える．この変数に対するある特定の値を観測したときに，どれだけの情報を受け取るかを考える
- 情報の量はxの値を得たときの「驚きの度合い」
- 起きそうにないこと事象が起きることを知れば，多くの情報量を得たと言える

ex)
「明日，太陽が東から昇ります」という情報と
「明日，雨が降ります」という情報はどちらが情報量が多いか？

情報量を測る尺度が必要

→情報量の定義へ

情報量

情報量をh(・)という関数で表す

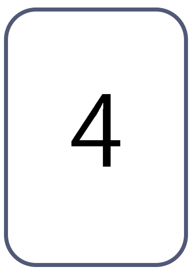
$$h(x,y)=h(x)+h(y)$$

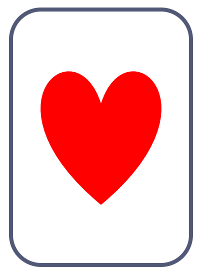
$$p(x,y)=p(x)\cdot p(y)$$

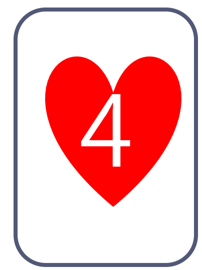
$$h(x)=-\log_2 p(x) \qquad (1.92)$$

単位：ビット

ex). トランプから1枚カードを抜くことを考える

x 

y 

x,y 

確率p(x) $\frac{1}{13}$　p(y) $\frac{1}{4}$　p(x,y) $\frac{1}{52}$

情報量h(x) $\log_2 13$　h(y) $\log_2 4$　h(x,y) $\log_2 52$

（情報論の）エントロピー (entropy)

- ある送信者が確率変数の値を受信者に送りたいとき，その操作で送られる情報の平均量は(1.92)を分布p(x)に関して期待値をとったものとなり，これをエントロピーという

$$H[x] = -\sum_{x} p(x)\log_2 p(x) \qquad (1.93)$$

熱力学に分子の無秩序さを表す「エントロピー（entropy）」という言葉があるが，上式とまったく同じ形をしている

エントロピーは情報の無秩序さ、あいまいさ、不確実さを表す尺度

ある事柄の発生確率がすべて同じとき すなわち何が起こるか予測がつかないときに最大で、発生確率の偏りが大きければ大きいほどエントロピーは小さくなる

情報論のエントロピー（２）

- 8個の取り得る変数 {a, b, c, d, e, f, g, h}
- それぞれの確率 {1/8, 1/8 , 1/8 , 1/8 , 1/8 , 1/8 , 1/8 , 1/8 }
- このときエントロピーは

$$H[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ ビット}$$

- 8個の取り得る変数 {a, b, c, d, e, f, g, h}
- それぞれの確率 {1/2, 1/4 , 1/8 , 1/16 , 1/64 , 1/64 , 1/64 , 1/64}
- このときエントロピーは

$$H[x] = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{1}{64}\log_2\frac{1}{64} = 2 \text{ ビット}$$

情報論のエントロピー（３）

- エントロピーの概念…確率変数の状態を規定するのに必要な情報量の平均量
- エントロピーの別の見方を考える

N個の同じ物体がたくさんの箱に分けられている状況を考える.
どの物体を最初のものとして選ぶかにN通りの場合があり,
次の物体はN-1通りあるので，N個の物体をどういう順序で入れるかはN!通りある
i番目の箱にはni!通りの順番付けがあるので，N個の物体の箱への入れ方の総数は

多重度
(multiplicity)

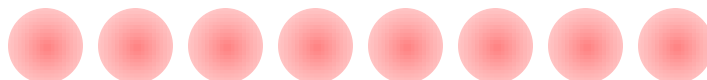
$$W = \frac{N!}{\prod_i n_i!} \qquad (1.94)$$

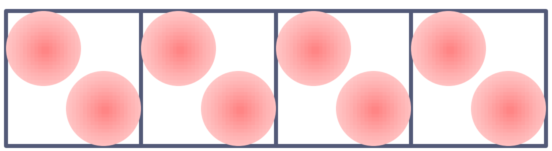
多重度の例

多重度 (multiplicity)

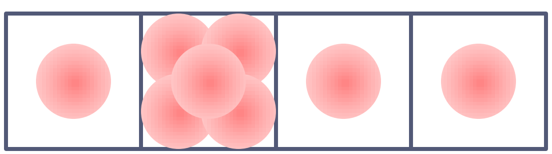
$$W = \frac{N!}{\prod_i n_i!} \qquad (1.94)$$

ex). N=8, 箱の数が4つのとき





$$W = \frac{8!}{2!2!2!2!} = 5040$$



$$W = \frac{8!}{1!5!1!1!} = 336$$

多重度からエントロピーへ

- 統計力学の観点からエントロピーを多重度から導出

$$H = \frac{1}{N}\ln W = \frac{1}{N}\ln N! - \frac{1}{N}\sum_i \ln n_i! \quad (1.95)$$

  - ここで，ni/Nを一定に保ったままN→∞という極限を考え，スターリングの近似式を用いると

$$\ln N! \approx N\ln N - N \quad (1.96)$$

多重度からエントロピーへ

- 統計力学の観点からエントロピーを多重度から導出

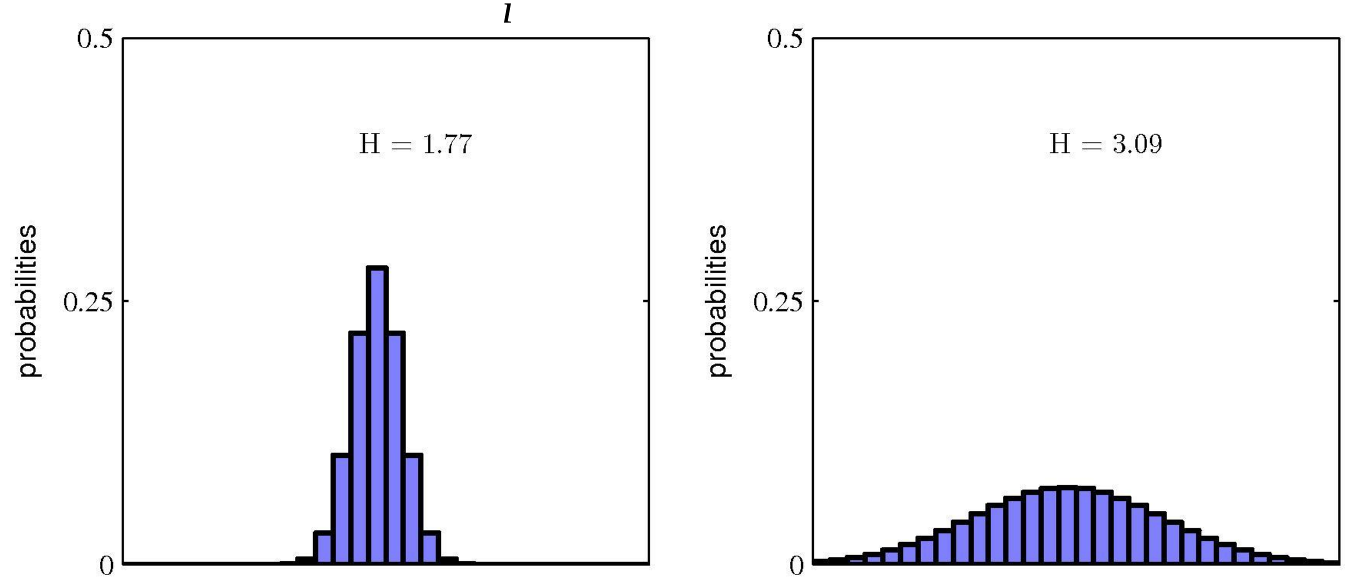
$$H=\frac{1}{N}\left\{(N\ln N-N)-\sum_i(n_i\ln n_i-n_i)\right\} \tag{1.97}$$

- ここで，p_iとは物体がi番目の箱に割り当てられる確率
- 物理学の用語で言えば，
- 箱の中の物体の特定の状態は
- ミクロ状態(微視的状態microstate)
- 比n_i/Nで表される物体の占有数の分布は
- マクロ状態(巨視的状態macrostate)
- 多重度Wはマクロ状態の重み(weight)とも呼ばれる

統計力学的エントロピーの定義

- 箱は離散確率変数Xの状態xiと解釈できるので，p(X=xi)=piとなる．すると確率変数Xのエントロピーは

$$H[p] = -\sum_i p(x_i) \ln p(x_i) \qquad (1.98)$$



- 広がりが大きい分布はエントロピーが大きい
- また，エントロピーは非負である

離散確率変数の最大エントロピー

- 最大のエントロピーを持つ確率分布
- Hを最大化するようにラグランジュ乗数法を用いて

$$\widetilde{H} = -\sum_i p(x_i)\ln p(x_i) + \lambda\left(\sum_i p(x_i) - 1\right) \qquad (1.99)$$

- ここから，Mをxiの状態の総数として

$$p(x_i) = \frac{1}{M}$$

$$H = \ln M$$

- となることがわかる（一様分布）

連続確率分布への拡張

- 連続変数xの分布p(x)に拡張する
- まず，xを等間隔の区間Δに分ける
- p(x)が連続であると仮定すれば，平均値の定理より

$$\int_{i\Delta}^{i+1\Delta} p(x)dx = p(x_i)\Delta \qquad (1.101)$$

- となるxiが必ず存在する
- i番目の区間に入る任意の値xに値xiを割り当てることで量子化を行うと，xiの値を観測する確率はp(xi)Δとなる
- ここから，離散分布のエントロピーは(1.101)より

$$H_\Delta = -\sum_i p(x_i)\Delta \ln(p(x_i)\Delta) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln\Delta \qquad (1.102)$$

微分エントロピー

$$H_\Delta = -\sum_i p(x_i)\Delta \ln(p(x_i)\Delta) = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln\Delta \quad (1.102)$$

- ここで(1.102)の右辺第二項をとりあえず無視して，Δ→0の極限を考えると

$$\lim_{\Delta\to 0}\left\{-\sum_i p(x_i)\Delta \ln p(x_i)\right\} = -\int p(x)\ln p(x)dx \qquad (1.103)$$

- 右辺の量は微分エントロピーを呼ばれている
- 離散と連続の場合のエントロピーはln Δだけ異なり
- この値はΔ→0で発散することがわかる
- これは連続変数を厳密に規定するのに無限のビット数が必要なことを表している

連続変数の場合のエントロピー最大化

- 離散変数では等確率が最大
- 連速変数では？以下の制約の下で最大化する

$$\int_{-\infty}^{\infty} p(x)dx = 1 \qquad (1.105)$$

$$\int_{-\infty}^{\infty} xp(x)dx = \mu \qquad (1.106)$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2 \qquad (1.107)$$

$$-\int_{-\infty}^{\infty} p(x)\ln p(x)dx + \lambda_1\left(\int_{-\infty}^{\infty} p(x)dx - 1\right)$$

$$+\lambda_2\left(\int_{-\infty}^{\infty} xp(x)dx - \mu\right) + \lambda_3\left(\int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx - \sigma^2\right)$$

連続変数の場合のエントロピー最大化

- 変分法により，この汎関数の微分を0とおいて

$$p(x) = \exp\left\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2\right\} \qquad (1.108)$$

- 最終的に

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \qquad (1.109)$$

- が得られる．これは正規分布である

$$H[x] = \frac{1}{2}\left\{1 + \ln(2\pi\sigma^2)\right\} \qquad (1.110)$$

- これは分散が大きくなればエントロピーが大きくなることも示している

相対エントロピーと相互情報量

- エントロピーをはじめとする情報理論のアイデアをパターン認識と関係づける
- 未知の分布p(x)があり，これを近似的にq(x)でモデル化したとする
- q(x)を用いてxの値を受信者に送るための符号化作業を構築したい
- 真の分布p(x)の代わりにq(x)を使うとxの値を特定するのに必要な追加(additional)情報量の平均はナットで測って

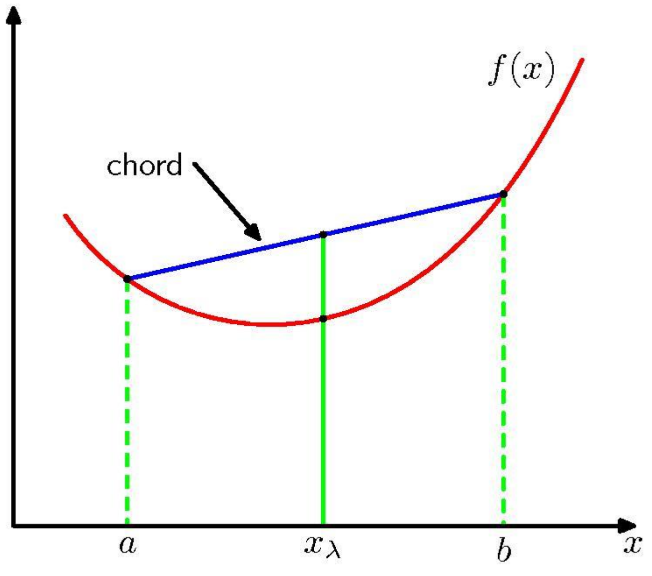
$$KL(p \| q) = -\int p(x) \ln q(x) dx - \left(-\int p(x) \ln p(x) dx\right)$$

$$= -\int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} dx \qquad (1.113)$$

- これは分布p(x)とq(x)の間の相対エントロピー(relative entropy)
- あるいはカルバック-ライブラーダイバージェンス(Kullback-Leibler divergence)あるいは略してKLダイバージェンスとして知られている

凸関数

- ここで，KLダイバージェンスは $KL(p \| q) \geq 0$ を満たし，なおかつ等式が成り立つのはp(x)=q(x)のとき，そのときに限ることを示す
- まず，凸関数(convex function)の概念を導入する
- 関数f(x)はすべての弦が関数に乗っているか，それよりも上にあるとき凸であるという



$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b) \qquad (1.114)$$

カルバック-ライブラーダイバージェンス

- 数学的帰納法を用いると(1.114)より凸関数f(x)が任意の点集合{xi}に対して,

$$f\left(\sum_{i=1}^{M}\lambda_i x_i\right) \le \sum_{i=1}^{M}\lambda_i f(x_i) \qquad (1.115)$$

- を満たすことができる．ここでΣλ=1である.
- (1.115)はイェンセンの不等式として知られている.
- λiを値xiを取る離散確率変数x上の確率分布として解釈すると

$$f(\mathbf{E}[x]) \le \mathrm{E}[f(x)] \qquad (1.116)$$

- と書ける．イェンセンの不等式をカルバック-ライブラーダイバージェンス(1.113)に適用することができ,

$$KL(p \| q) = -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}dx \ge -\ln\int q(x)dx = 0 \qquad (1.118)$$

- が得られる

KLダイバージェンスと密度推定（１）

- データ圧縮と密度推定（未知の確率分布のモデル化の問題）は密接に関係しており，最も効率的な圧縮は真の分布を知っているときに達成される
- 真の分布と異なる分布を使えば，非効率な符号化となり，送信しなければならない追加情報量は平均して2つの分布の間のカルバック-ライブラーダイバージェンスと等しくなる
- データが未知の分布p(x)から生成されるとき，それをモデル化してみよう
- 可変なパラメータθをもつパラメトリックな分布q(x|θ)（たとえば多変量正規分布）を使って近似することを考える
- θを決める1つの方法はp(x)とq(x|θ)の間のカルバック-ライブラーダイバージェンスをθについて最小化することが考えられる
- しかし，p(x)を知らないのでこれを直接行うことはできない…

KLダイバージェンスと密度推定（２）

- p(x)から得られた有限個の訓練点の集合x_n\{n=1,⋯..,N\}が手元にある
- p(x)に関する期待値はこれらの点での有限和に近似でき，

$$KL(p \| q) \approx \frac{1}{N} \sum_{n=1}^{N} \left\{ -\ln q(x_n \mid \theta) + \ln p(x_n) \right\} \qquad (1.119)$$

- となる．右辺第二項はθとは独立であり，最初の項は訓練集合を使って評価した分布の下でのθの負の対数尤度である．
- つまり，KLダイバージェンスの最小化は尤度の最大化と等価である

相互情報量

- 2つの変数集合xとyの同時分布p(x, y)を考える
- 変数の集合が独立であれば同時分布は周辺分布の積に分解され p(x, y)=p(x)p(y)となる
- 変数が独立でなければ，独立に近いかどうかを知るために，同時分布と周辺分布の積の間のKLダイバージェンスを考えることができ，

$$
\begin{aligned}
I[x,y] &\equiv KL(p(x,y) \| p(x)p(y)) \\
&= -\iint p(x,y) \ln\left(\frac{p(x)p(y)}{p(x,y)}\right) dxdy
\end{aligned}
\qquad (1.120)
$$

- これは変数x,yの間の相互情報量(mutual information)と呼ばれる
- 相互情報量はyの値を知ることによってxに関する不確実性がどれだけ減少するかを表す.
- ベイズ的に言えばp(x)をxの事前分布，p(x|y)は新たなデータyを観測した後の事後分布と考えられる．したがって，新たにyを観測した結果として，xに関する不確実性が減少した度合いを表している