Deep Learning 輪読会 2017

第3章　確率と情報理論

2017.11.06

東京大学工学部 松尾研究室

B4 松嶋 達也 (@__tmats__)

構成

- はじめに
- 3.1 なぜ確率なのか
- 3.2 確率変数
- 3.3 確率分布
- 3.4 周辺確率
- 3.5 条件付き確率
- 3.6 条件付き確率の連鎖律
- 3.7 独立と条件付き確率
- 3.8 期待値，分散と共分散
- 3.9 一般的な確率分布
- 3.10 一般的な関数の有用な性質
- 3.11 ベイズ則
- 3.12 連続変数の技術的詳細
- 3.13 情報理論
- 3.14 構造化確率モデル

はじめに

- 前章の議論：深層学習の理解に必要な線形代数の解説

- 本章の議論：確率と情報理論に関する解説
  - 確率論は不確実な命題を表現する数学的な枠組み
  - 人工知能の応用での確率論の利用
    - 確率法則によってAIシステムがどのように推論するべきかを知って，確率論を使うことで得られる多様な表現の計算や近似のためのアルゴリズムを設計する
    - 確率と統計を使って，提案されたAIシステムの振る舞いを理論的に解析する

3.1 なぜ確率なのか

- 不確実性が存在する中での推論能力が要求される
  - 確率を使って不確実性を定量化

- 不確実性を生み出す可能性のある要因
  - モデル化されるシステムに固有の確率性
    - Ex) 素粒子の力学
  - 不完全な可観測性
    - システムの振る舞いを決める変数の全てを観測できない場合
      - Ex)モンティ・ホール問題
  - 不完全なモデリング
    - 観測した情報を破棄しなければならないモデルを使う時
      - Ex) 離散化

3.1 なぜ確率なのか

- **頻度確率** (frequent probability)
  - 事象が起こる割合に直接関係しているもの
    - 繰り返すことができない命題にはそのまま適用できない
- **ベイズ確率** (Bayesian probability)
  - **信念の度合い**(degree of belief)を表現するもの

- ベイズ確率が頻度確率とまったく同じよう振る舞うとみなすと．不確実性についての常識的な推論ができる
  - Ramsey, F. P. (1926). Truth and probability.

- 確率論は，不確実性を扱うための論理学の拡張とみなせる

3.2 確率変数

- **確率変数** (random variable)
  - 無作為に異なる値をとることができる変数
    - 通常，確率変数自体は小文字の単純な書体で，変数としてとることのできる値は小文字の筆記体で表記する
      - Ex) x_1, x_2: 確率変数xが取りうる値
      - Ex) $\boldsymbol{x}$: 確率変数**x**の値の1つ(ベクトル値の変数)
    - 離散値でも連続値でもよい

3.3 確率分布

- **確率分布** (probability distribution)
  - 確率変数や確率変数の集合が取りうる値それぞれの尤もらしさを記述するもの
    - 変数が離散か連続かで記述する方法が決まる
      - 離散変数のとき：確率質量関数
      - 連続変数のとき：確率密度関数

3.3 確率分布 – 離散変数と確率質量関数

- **確率質量関数** (probability mass function, PMF)
  - 離散変数の確率分布
  - ある確率変数の状態から，その確率変数がその状態をとる確率への写像

- 確率質量関数Pが満たすべき性質
  - Pの定義域は，xが取りうる状態全ての集合でなければならない
  - 発生しない事象の確率は0であり，これよりも発生確率が低くなる状態はない．同様に，発生することが保証されている事象の確率は1であり，これよりも発生確率が高くなる状態はない.
    - $\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1$
  - 正規化
    - $\sum_{x \in \mathrm{x}} P(x) = 1$

3.3 確率分布 – 離散変数と確率質量関数

- **同時確率分布** (joint probability distribution)
  - 多変数の確率分布
  - $P(\mathrm{x} = x, \mathrm{y} = y)$ は同時に$\mathrm{x} = x$かつ$\mathrm{y} = y$である確率を表し，簡潔に$P(x, y)$とも書ける．

3.3 確率分布 – 連続変数と確率密度関数

- **確率密度関数** (probability density function, PDF)
  - 連続変数の確率分布

- 確率密度関数pが満たすべき性質
  - pの定義域は，xが取りうる状態全ての集合でなければならない
  - $\forall x \in \mathrm{x}, p(x) \geq 0$，ただし，$\mathrm{p(x)} \leq 1$は必要条件ではないことに注意
  - $\int P(x)dx = 1$

- 確率密度関数$p(x)$からは特定の状態の確率は直接的に得られない
  - 容積がδxの微小領域にある確率は$p(x)\delta x$で与えられる

3.4 周辺分布

- **確率周辺分布**(marginal probability distribution)
  - 変数の集合の確率分布が分かっているとき，その部分集合の確率分布
    - Ex) 離散変数x, y について，$P(\mathrm{x}, \mathrm{y})$が分かっているとすると，$P(\mathrm{x})$は**確率の加法定理**(sum rule)で求められる

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y).$$

    - 連続変数の場合

$$p(x) = \int p(x, y) dy.$$

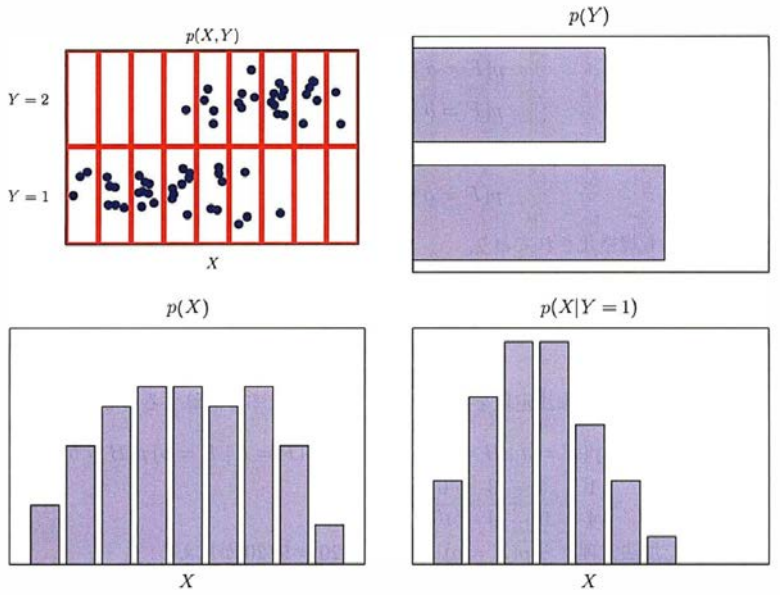


図 1.11 9 つの値を取り得る変数 X と 2 つの値を取り得る Y の 2 変数に対する分布の図．左上の図はこれらの変数の同時分布から生成した 60 個のサンプル点を示す．残りの図は周辺分布 $p(X), p(Y)$ と，左上の図の下側の行に対応する条件付き分布 $p(X \mid Y = 1)$ のヒストグラム推定である．

出典:パターン認識と機械学習 上

3.5 条件付き確率

- **条件付き確率**(conditional probability)
  - ある事象が起きたという条件の下で，別な事象が起きる確率

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$$

  - $P(x = \mathrm{x}) > 0$のときのみ定義される

3.6 条件付き確率の連鎖律

- **確率の連鎖律**(chain rule)または**確率の乗法定理**(product rule)
  - 多数の確率変数における同時確率分布は，たった一つの変数に対する条件付き確率分布に分解できる場合がある

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)}).$$

- Ex)
$$P(\mathrm{a}, \mathrm{b}, \mathrm{c}) = P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b}, \mathrm{c})$$
$$P(\mathrm{b}, \mathrm{c}) = P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c})$$
$$P(\mathrm{a}, \mathrm{b}, \mathrm{c}) = P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}).$$

3.7 独立と条件付き分布

- 2つの確率変数xとyの確率分布が，xだけを含むものとyだけを含むものの2つの因子の積で表現できるならば，この2つの確率変数は**独立**(independent)である.
  - $\mathrm{x} \perp \mathrm{y}$と表記

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y)$$

- 確率変数zが与えられた下で，2つの確率変数xとyの条件付き確率分布が，zの全ての値において上記の方法で因数分解されるならば，xとyは**条件付き独立**(conditionally independent)である.
  - $\mathrm{x} \perp \mathrm{y} | \mathrm{z}$と表記

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x} = x, \mathrm{y} = y \mid \mathrm{z} = z) = p(\mathrm{x} = x \mid \mathrm{z} = z)p(\mathrm{y} = y \mid \mathrm{z} = z)$$

3.8 期待値，分散と共分散 – 期待値

- 確率分布$P(\mathrm{x})$に関する関数$f(x)$の**期待値**(expectation, expected value)
  - Pからxが抽出された下で，fがとる値の平均または平均値
    - 離散変数の場合

$$\mathbb{E}_{\mathrm{x}\sim P}[f(x)] = \sum_x P(x)f(x)$$

    - 連続変数の場合

$$\mathbb{E}_{\mathrm{x}\sim p}[f(x)] = \int p(x)f(x)dx$$

3.8 期待値，分散と共分散 – 分散

- **分散**(variance)
  - 確率分布から様々なxの値を抽出した場合，その確率変数xの関数値のばらつきの度合いを示す指標

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

- **標準偏差**(standard deviation)
  - 分散の平方根

3.8 期待値，分散と共分散 – 共分散

- **共分散**(covariance)
  - 2つの変数の大きさと共に，それらの値がどの程度線形的に互いに関連しているか

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])\left(g(y) - \mathbb{E}\left[g(y)\right]\right)\right]$$

- **共分散行列**(covariance matrix)

$$\mathrm{Cov}(\mathbf{x})_{i,j} = \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j).$$

  - 共分散行列の対角成分は分散

$$\mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_i) = \mathrm{Var}(\mathrm{x}_i)$$

- **相関**(correlation)
  - 各変数の寄与度を正規化
    - 変数の大きさの影響を排除して，変数の関連度合いのみを測る

3.9 一般的な確率分布 – ベルヌーイ分布

- **ベルヌーイ分布**(Bernoulli distribution)
  - 1つの2値の確率変数における分布
  - 1つのパラメータ$\phi \in [0,1]$で制御される
    - ベルヌーイ分布の性質

$$P(\mathrm{x} = 1) = \phi$$

$$P(\mathrm{x} = 0) = 1 - \phi$$

$$P(\mathrm{x} = x) = \phi^{x}(1 - \phi)^{1-x}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi$$

$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1 - \phi)$$

3.9 一般的な確率分布 – マルチヌーイ分布

- **マルチヌーイ**(multinoulli)または**カテゴリ**(categorical)**分布**
  - k個の異なる状態をとる1つの離散変数における分布(但し，kは有限の値)
    - ベクトル$\boldsymbol{p} \in [0,1]^{k-1}$でパラメータ化 ($p_i$は$i$番目の状態の確率)
    - 最後のk番目の状態の確率は$1 - \mathbf{1}^{\top}\boldsymbol{p}$で求められる
    - 対象カテゴリの分布を参照
  - **多項分布**(multinomial distribution)の特別な形
    - $\{0, \ldots, n]^k$に含まれるベクトルの分布
    - マルチヌーイ分布からn個のサンプルが取られたときに，k個の各カテゴリが選ばれた回数をあらわす

- ベルヌーイ分布とマルチヌーイ分布は，離散変数をモデル化

3.9 一般的な確率分布 – ガウス分布

- **ガウス分布**(Gaussian distribution)または**正規分布**(normal distribution)

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

  - 2つのパラメータ$\mu \in \mathbb{R}$と$\sigma \in (0,\infty)$により決定

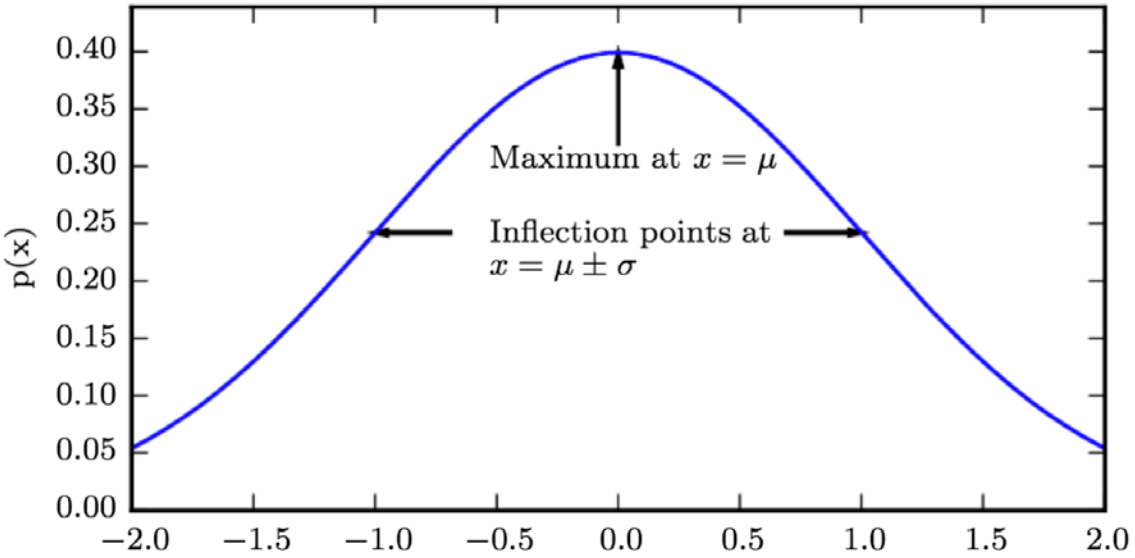


Figure 3.1: The normal distribution. The normal distribution $\mathcal{N}(x;\mu,\sigma^2)$ exhibits a classic "bell curve" shape, with the x coordinate of its central peak given by μ, and the width of its peak controlled by σ. In this example, we depict the **standard normal distribution**, with $\mu = 0$ and $\sigma = 1$.

3.9 一般的な確率分布 – ガウス分布

- 分布の**精度**(precision)$\beta = \sigma^{-2}$を利用した表現
  - 確率密度関数を頻繁に評価する必要がある場合に効率的

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

3.9 一般的な確率分布 – ガウス分布

- 正規分布は，選択すべき実数に対する分布の形式について事前知識がない場合，最初の選択として正しい
  - **中心極限定理**(central limit theorem)
    - 多くの独立な確率変数の和が近似的に正規分布になる
      - 複雑な系の多くでは，たとえその系がより構造化された振る舞いをする部分に分解できたとしても，正規分布に従う雑音としてうまくモデル化される
  - 同じ分散を持つ全ての確率分布の中で，正規分布は実数における不確実性の最大となる量を符号化する
    - PRML1.6節(p.53)参照

3.9 一般的な確率分布 – ガウス分布

- **多変量正規分布**(multivariable normal distribution)
  - 正規分布の$\mathbb{R}^n$への一般化

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

  - パラメータ$\boldsymbol{\mu}$(分布の平均)と$\boldsymbol{\Sigma}$(分布の共分散行列)により決定
  - **精度行列**(precision matrix)$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}$を使った表現

$$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\beta}^{-1}) = \sqrt{\frac{\det(\boldsymbol{\beta})}{(2\pi)^n}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\beta}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

3.9 一般的な確率分布 – ガウス分布

- **多変量正規分布**(multivariable normal distribution)
  - 共分散行列を対角行列に限定することが多い.
    - 単位行列のスカラー倍に限定したものは**等方性**(isotropic)**ガウス分布**

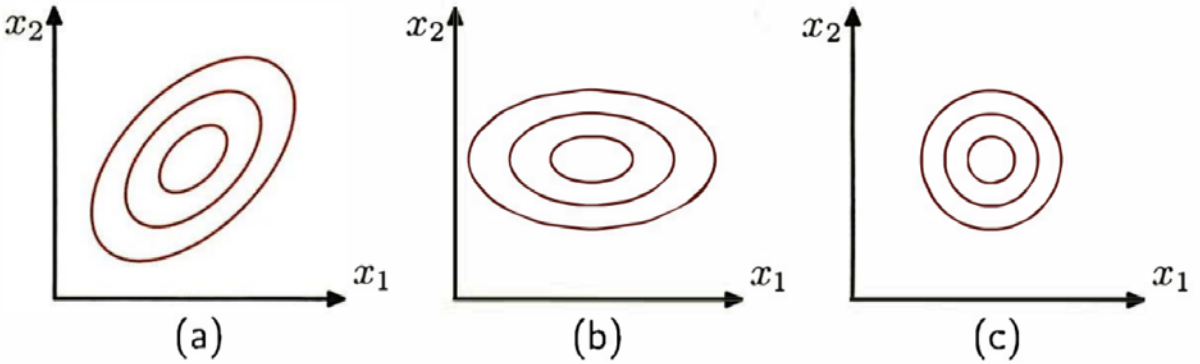


図 2.8 2 次元空間のガウス分布の確率密度が一定になる等高線を示したもの．それぞれ，共分散行列が (a) 一般のもの，(b) 対角，すなわち，等高線の楕円が座標軸に沿っているもの，および (c) 単位行列に比例する，すなわち，等高線が同心円になっているものである．

出典:パターン認識と機械学習 上

3.9 一般的な確率分布 – 指数分布とラプラス分布

- **指数分布**(exponential distribution)
  - 深層学習の観点では$x = 0$で尖った部分を持つ確率分布が必要になることが多い

$$p(x; \lambda) = \lambda \mathbf{1}_{x \geq 0} \exp(-\lambda x)$$

- **ラプラス分布**(Laplace distribution)
  - 任意の点μで確率質量の尖った峰を作ることのできる，指数分布に密接に関連した確率分布

$$\text{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x - \mu|}{\gamma}\right)$$

3.9 一般的な確率分布 – ディラック分布と経験分布

- **ディラック分布**(Dirac distribution)
  - **ディラックのデルタ関数**(Dirac delta function)$\delta(x)$を使って確率密度関数を定義
    - 0以外のところは全て0だが，積分すると1になるように定義される関数

$$p(x) = \delta(x - \mu)$$

  - **経験分布**(empirical distribution)の構成要素としてよく用いられる
    - データ集合または事例の集合を形成するm個の点それぞれで確率質量の値が$\frac{1}{m}$
    - 連続変数の経験分布を定義する場合にのみ必要(離散変数の時マルチヌーイ分布)
    - 訓練データの尤度を最大化する確率密度(5.5節参照)

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\delta(\boldsymbol{x} - \boldsymbol{x}^{(i)})$$

3.9 一般的な確率分布 – 分布の混合

- **混合分布**(mixture distribution)
  - 単純な他の確率分布を組み合わせて定義した確率分布
    - $P(c)$は成分の性質のマルチヌーイ分布

$$P(\mathrm{x}) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} \mid \mathrm{c} = i)$$

    - Ex) 実数値の変数の経験分布
      - 訓練事例それぞれにディラック分布の成分を1つもつ混合分布

- **潜在変数**(latent variable)
  - 直接には観測できない確率変数
    - Ex) 混合モデルの要素情報の変数c

3.9 一般的な確率分布 – 分布の混合

- **混合ガウス**(Gaussian mixture)モデル
  - 構成要素$p(\mathbf{x}|c=i)$がガウス分布
  - 各要素には，別々にパラメータ化された平均$\boldsymbol{\mu}^{(i)}$と共分散$\boldsymbol{\Sigma}^{(i)}$が存在
    - 共分散行列に制約を課すことがある
  - 混合ガウスモデルのパラメータは各要素iに対して**事前確率**(prior probability)を規定する
    - 事前確率：$\alpha_i = P(c=i)$
    - 「事前」…$\mathbf{x}$が観測される前の，cに関するモデルの信念
      - Cf) **事後確率**(posterior probability)$P(c|\boldsymbol{x})$… $\mathbf{x}$が観測された後に計算される
  - 十分な数の混合ガウスモデルを使えば，どんな滑らかな密度も，ゼロでない数の誤差で近似できるという意味で，**密度の万能近似器**(universal approximator)である

3.10 一般的な関数の有用な性質

- **ロジスティックシグモイド**(logistic sigmoid)
  - 値域が(0,1)でϕパラメータの有効な範囲内にある
    - ベルヌーイ分布のϕパラメータを生成する際によく使われる

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

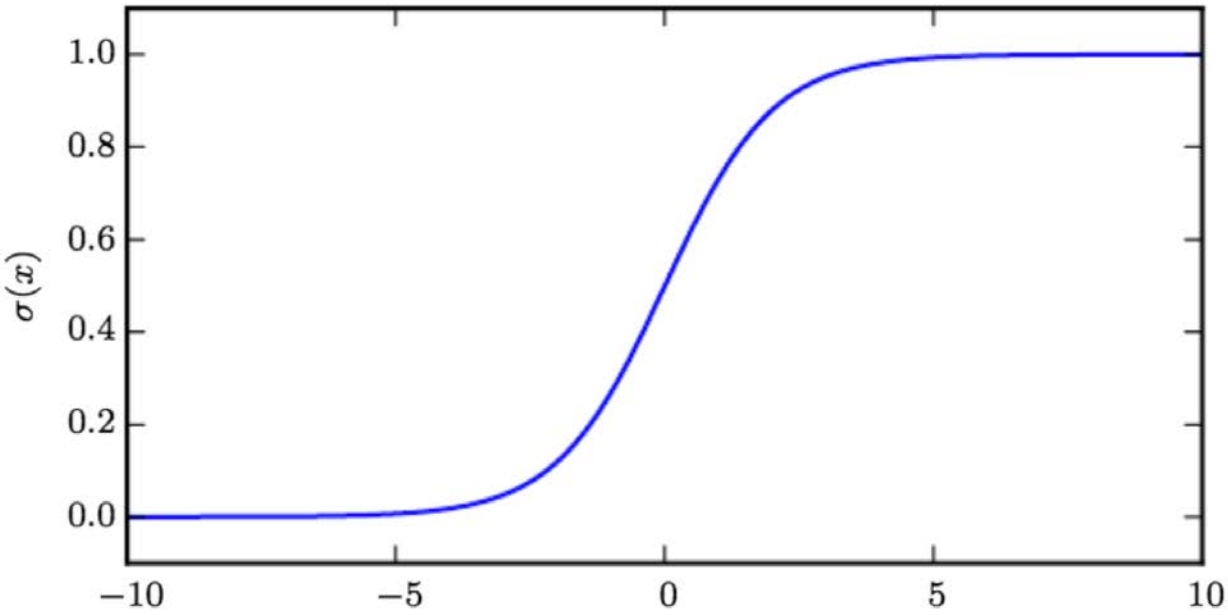


Figure 3.3: The logistic sigmoid function.

3.10 一般的な関数の有用な性質

- **ソフトプラス**(softplus)関数
  - 値域が(0,∞)
    - 正規分布のβ, σパラメータを生成する際によく使われる

$$\zeta(x) = \log\left(1 + \exp(x)\right)$$

  - **正の部関数**(positive part function) $x^+ = \max\{0, x\}$の平滑化を意図したもの

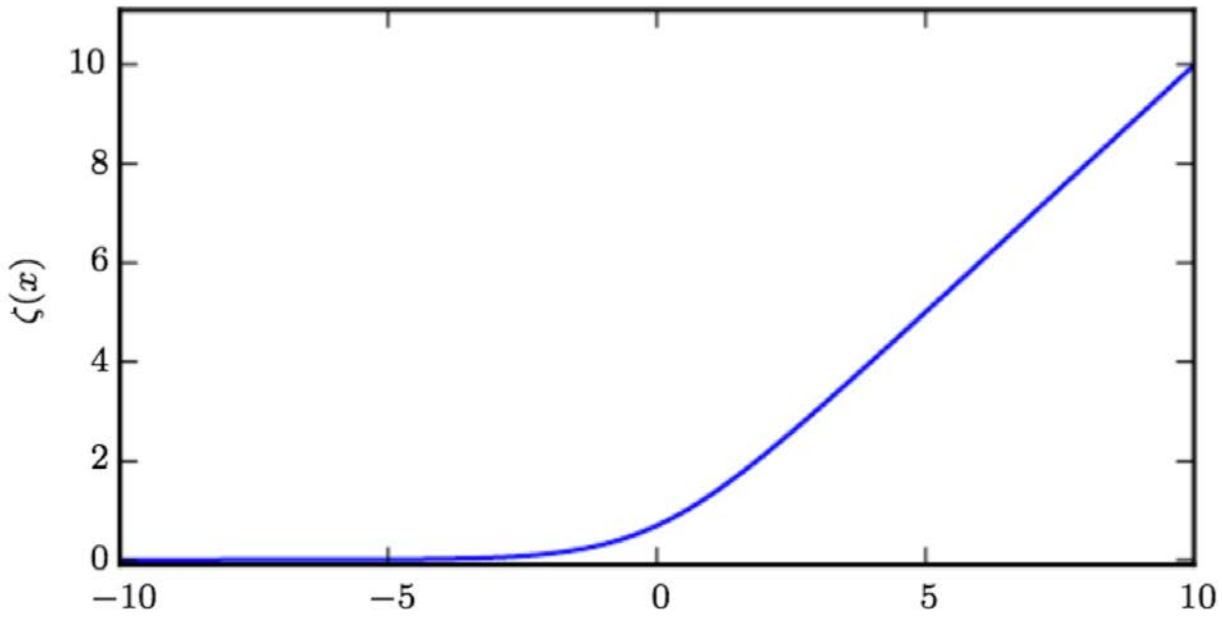


Figure 3.4: The softplus function

3.11 ベイズ則

- **ベイズ則**(Bayes' rule)
  - $P(\mathrm{y}|\mathrm{x})$が分かっていて，$P(\mathrm{x})$が既知ならば，$P(\mathrm{x}|\mathrm{y})$を算出できる

$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}.$$

    - 但し，$P(\mathrm{y}) = \sum_x P(y|x)P(x)$であり，通常計算できる

3.12 連続変数の技術的詳細 – 測度論

- **測度論**(measure theory)
  - 矛盾を避けながら，確率を計算できる集合族を特徴付ける
  - **測度零**(measure zero)
    - 点の集合が無視できるぐらい小さい
  - **ほとんど至るところで**(almost everywhere)
    - ほとんど至るところで見られる性質は，測度零の集合を除いた全空間で見られる
    - 例外が占める空間は無視できるため，多くの応用で．それを無視しても問題ない

3.12 連続変数の技術的詳細 – 確率変数の変数変換

- 確率変数の変数変換
  - お互いの決定論的関数となる連続確率変数の扱い
    - Ex) 2つの確率変数$\mathbf{x}, \mathbf{y}$があり，可逆で連続かつ微分可能な変換gを使って$\boldsymbol{y} = g(\mathbf{x})$が成立すると仮定

$$p_x(\boldsymbol{x}) = p_y(g(\boldsymbol{x})) \left| \det \left(\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}} \right) \right|$$

  - 変換された変数の同時確率密度関数に**ヤコビ行列**(Jacobian matrix)の行列式がかかる
    - ヤコビ行列

$$—J_{i,j} = \frac{\partial x_i}{\partial y_j}$$

3.13 情報理論 – 自己情報量，シャノンエントロピー

- 情報の量的な表現
- 事象$\mathrm{x} = x$の**自己情報量**(self-information)
  - $\mathrm{I}(\mathrm{x})$の定義は**ナット**(nats)の単位で書かれる
    - 1ナットは確率$\frac{1}{e}$の事象を観測したときに得られる情報量

$$I(x) = -\log P(x)$$

- **シャノンエントロピー**(Shannon entropy)
  - 確率密度全体の不確実性を量的に表現
  - その分布から抽出される事象に期待される情報量

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]$$

  - ほぼ決定論的な分布のエントロピーは低く，一様分布に近い分布のエントロピーは高い
  - xが連続であるとき，**微分エントロピー**(differential entropy)と呼ばれる

3.13 情報理論 - KLダイバージェンス

- **カルバック・ライブラーダイバージェンス**(Kullback-Leibler(KL) divergence)
  - 同じ確率変数xに対して，異なる確率分布P(x)とQ(x)があるとき，この2つの分布にどれだけの差があるのかを測る

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right]$$

  - KLダイバージェンスの性質
    - 非負
    - 離散変数において，同じ分布である場合，連続変数において，分布が「ほとんど至るところで」等しくなる場合に限り，0となる
    - 非対称

$$D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$$

3.13 情報理論 – 交差エントロピー

- **交差エントロピー**(cross-entropy)

$$H(P,Q) = H(P) + D_{\mathrm{KL}}(P\|Q)$$

  - KLダイバージェンスの左側の項が削除されている

$$H(P,Q) = -\mathbb{E}_{\mathrm{x}\sim P} \log Q(x)$$

    - Qに関する交差エントロピーの最小化は，KLダイバージェンスの最小化と等しい

3.14 構造化確率モデル

- 1つの関数で同時確率分布の全体を記述することは効率が悪い
  - 掛け合わせが可能なたくさんの因子に確率分布を分割する
    - Ex) 3つの確率変数a,b,cがあるとする．aはbの値に影響を与え，bはcの値に影響を与えるが，bが与えられた下でaとcは独立

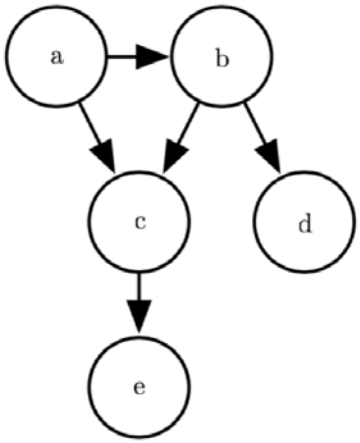
$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{b}).$$

  - グラフを使ってこの因数分解を表現できる
    - **構造化確率モデル**(structured probabilistic model)，あるいは**グラフィカルモデル**(graphical model)と呼ぶ
      - 確率分布を記述するための表現手段として，有向グラフ，無向グラフが存在

3.14 構造化確率モデル

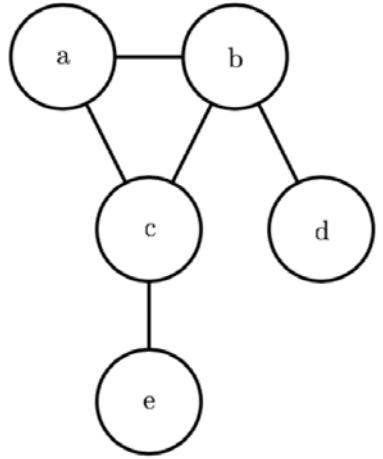
- **有向**(Directed)モデル
  - 向きのある辺を使ったグラフ
  - 条件付き確率分布への因数分解で表現

$$p(\mathbf{x}) = \prod_i p\left(\mathrm{x}_i \mid Pa_{\mathcal{G}}(\mathrm{x}_i)\right).$$



$$p(\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e}) = p(\mathrm{a})p(\mathrm{b}\mid\mathrm{a})p(\mathrm{c}\mid\mathrm{a},\mathrm{b})p(\mathrm{d}\mid\mathrm{b})p(\mathrm{e}\mid\mathrm{c})$$

- **無向**(Undirected)グラフ
  - 向きがない辺を使ったグラフ
  - 関数の集合への因数分解で表現
  - 因子は単に関数であり，確率分布ではない
    - Φ関数の積の全状態の総和か積分で定義される正規化定数Zで割って，確率分布を得る



$$p(\mathrm{a},\mathrm{b},\mathrm{c},\mathrm{d},\mathrm{e}) = \frac{1}{Z}\phi^{(1)}(\mathrm{a},\mathrm{b},\mathrm{c})\phi^{(2)}(\mathrm{b},\mathrm{d})\phi^{(3)}(\mathrm{c},\mathrm{e})$$

$$p(\mathbf{x}) = \frac{1}{Z}\prod_i \phi^{(i)}\left(\mathcal{C}^{(i)}\right).$$

参考文献 1

- パターン認識と機械学習 上
  - C.M. ビショップ (著), 元田 浩 (監訳), 栗田 多喜夫 (監訳), 樋口 知之 (監訳), 松本 裕治 (監訳), 村田 昇 (監訳)
- 東京大学工学教程 基礎系 数学 確率・統計I
  - 縄田 和満

参考文献 2

- Deep Learning
  - Ian Goodfellow, Yoshua Bengio, Aaron Courville
  - 日本語版
    https://www.amazon.co.jp/%E6%B7%B1%E5%B1%A4%E5%AD%A6%E7%BF%92-Ian-Goodfellow/dp/4048930621