Modelling of Priors

For Bayesian Classification

Contents

- Modelling of the Prior
- Priors from experiments
- Uninformed Priors
- Bayesian Classification: Discussion

Modelling of Prior

- So far, we have discussed how to determine the likelihood $p(\mathbf{x}|C)$ (training)
- Now, it needs to discuss how to determine the prior $p(C)$
- To determine posterior $p(C|\mathbf{x})$ with the help of the theorem of Bayes, in a form of synthetic data sets by sampling from the joint distribution $\mathrm{p}(\mathbf{x},\mathrm{C}) = p(\mathbf{x}|C) \cdot p(C) \rightarrow$ Generative Classifiers
- Possible origins of priors:
  1) **From experiments,** e.g. in the case of sequential data: the prior for the classification at time t depends on the state at time t-1
  2) **"Uninformed" / subjective:** from prior knowledge (... from whichever source)

Priors from Experiments

- Requirement: the prior distribution should have the same algebraic form as the likelihood function → "Conjugate Prior"
- Example: Estimation of the parameter μ of a Bernoulli distribution with $p(x) = \mu^x \cdot (1-\mu)^{(1-x)}$
  - N experiments
    - in n_+ cases the result is "1"
    - in n_- cases the result is "0"
    - $n_+ + n_- = N$

→ Maximum Likelihood estimation: $\mu = n_+ / N$
Can lead to overfitting → prior for μ?

Priors from Experiments

- Bayesian estimation of μ: $p(\mu \mid n_+) \propto p(n_+ \mid \mu) \cdot p(\mu)$
- $p(n_+ \mid \mu)$ follows a binomial distribution :

$$p(n_+ \mid \mu) = \frac{N!}{n_+! \cdot (N - n_+)!} \mu^{n_+} \cdot (1 - \mu)^{N - n_+}$$

- Priori distribution for μ?
  - Conjugate prior: Beta distribution with hyperparameters a, b:

$$p(\mu) = p(\mu \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)} \cdot \mu^{a-1} \cdot (1 - \mu)^{b-1}$$

- Resulting posterior:

$$p(\mu \mid n_+) \propto p(n_+ \mid \mu) \cdot p(\mu) \propto \mu^{n_+ + a - 1} \cdot (1 - \mu)^{N - n_+ + b - 1}$$

Priors from Experiments

- Resulting posterior:

$$p(\mu \mid n_+) \propto p(n_+ \mid \mu) \cdot p(\mu) \propto \mu^{n_+ + a - 1} \cdot (1-\mu)^{N - n_+ + b - 1}$$

- Interpretation:
  - ➢ a -1 ... The number of trials with x = 1 from “earlier experiments“, which formed the basis of the prior
  - ➢ b -1 ... The number of trials with x = 0 from “earlier experiments“, which formed the basis of the prior
- Simplifies the processing of sequential data

Priors from Experiments

- Conjugate priors for other distributions :

Likelihood	Parameter	Conjugate prior	Hyper-parameter	Posterior parameter
Binomial	μ	Beta	a,b	$a+n_+$, $b+(N-n_+)$
Multinomial	$\boldsymbol{\mu}$ ($\Sigma\mu_i = 1$)	Dirichlet	$\mathbf{a}$	a_i+n_{i+}
Normal, σ known	$\boldsymbol{\mu}$	Normal	$\mu_0,\ \sigma_0^2$	$\dfrac{\mu_0/\sigma_0^2 + \sum x_i/\sigma^2}{1/\sigma_0^2 + 1/\sigma^2}$
Normal, μ known	w (Precision)	Gamma	$\alpha,\ \beta$	$\alpha+n/2$, $\beta+1/2\ \Sigma(x_i-\mu)^2$

Uninformed Priors

- A priori probabilities from minimal additional information
- Subjective priors (without measurements / experiments)

  → Principle of Maximum Entropy (ME):

$$p_{ME} = \operatorname{argmax}_p \int_x -p(x)\log_2 p(x)\,dx$$

- Prior knowledge concerning the value range or moments of the distribution can be used to formulate of constraints for p_{ME}

Uninformed Priors

- Example for ME-Priors:
  - Known value range with $a \leq x \leq b$: $\int_{x=a}^{b} p(x)dx = 1$

    → Uniform distribution in the interval (a,b)

    → also applies for $(-\infty, +\infty)$ → in this case: ML-classification!
  - Known expected value m, $x \geq 0$: $\int_{x} x \cdot p(x)dx = m$

    → Exponential distribution: $p(x) = \frac{1}{m} \cdot e^{-\frac{x}{m}}$
  - Known expected value m, known variance s^2:

$$\int_{x} x \cdot p(x)dx = m \qquad \int_{x} (x-m)^2 \cdot p(x)dx = s^2$$

→ Normal distribution $N(m,s^2)$

Bayesian Classification: Discussion

- Bayesian classification (and extensions) has many applications
- There are many variants depending on the models used for the individual components
- Bayesian classification delivers optimal results **if**
  - The assumptions about the likelihood function and the priors are correct
  - The training data are representative for the classes
  - There are enough training data to estimate the parameters of the models reliably
- Problems occur when one of these assumptions is not justified

Bayesian Classification: Discussion

- Examples of problems:
  - Assumption: the assumptions about the likelihood function and the prioris are correct

    → Possible problem: unknown / wrong number of clusters for one or more classes in feature space
  - Assumption: The training data are representative

    → Possible problem: training data only for objects in the sun, not for objects in the shadow
  - Assumption: There are enough training data

    → Posible problem: not enough training data

    → reliable determination of the parameters may be impossible

Bayesian Classification: Discussion

- There is no mechanism to take into account uncertainties in the probabilities

➢ If the requirements are not fulfilled, Bayesisan classification may yield suboptimal results

- How to describe the quality of the results?

- How to determine the priors?

- Modelling the distribution of the data may require more parameters and, therefore, more training data than direct models of the posterior distribution

➢ If the requirements are not fulfilled, Bayesisan classification may yield suboptimal results