National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
Institute of Physics and Technology

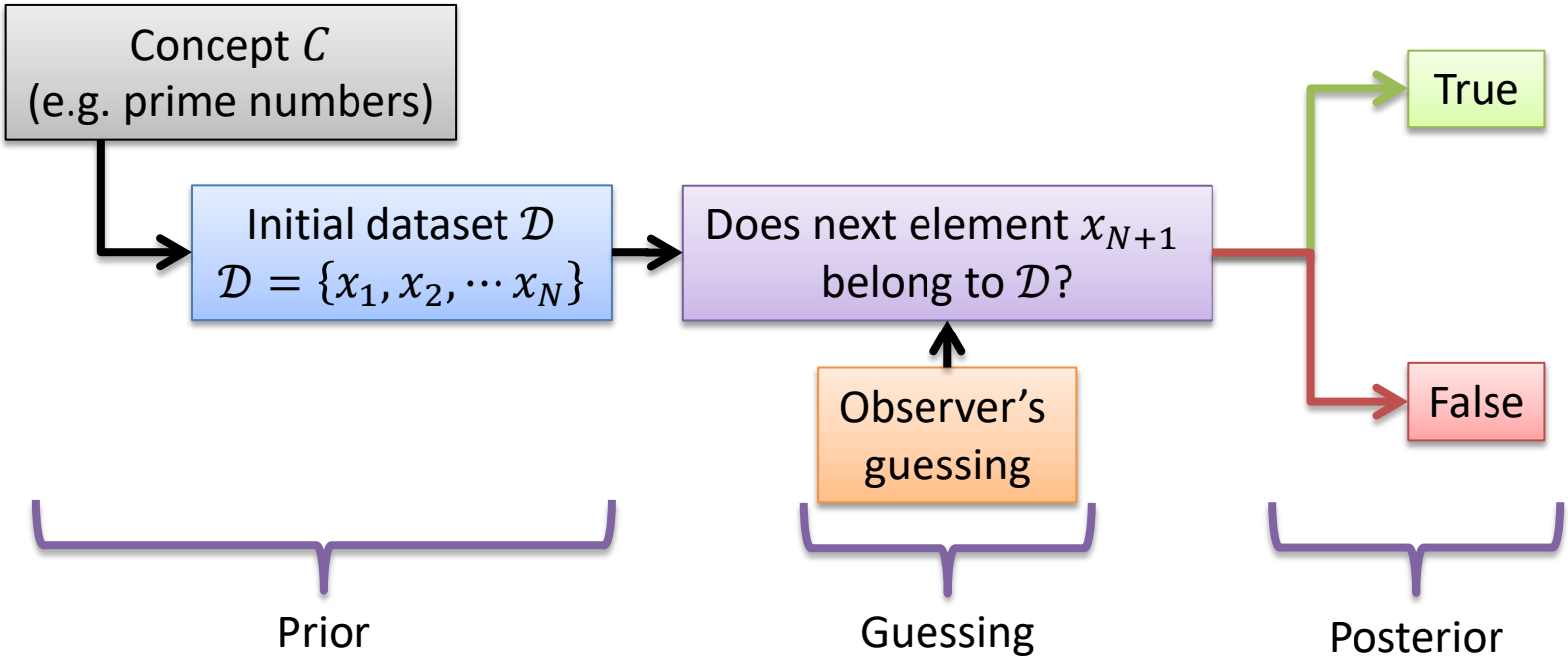
Lecture 3
Generative Models for Real-World Systems. Bayesian Concept

Dmytro Progonov,
PhD, Associate Professor,
Department Of Physics and Information Security Systems

Content

- Bayesian concept learning;
- Prior and posterior processing;
- Naïve Bayes classifier;
- Bayesian approach applications.

Bayesian concept learning



$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) \propto p(\mathbf{x}|y = c, \boldsymbol{\theta}) \cdot p(y = c|\boldsymbol{\theta})$$

- Observer's answer
- Supposed class
- Observed features
- Unknown parameters

Prior and posterior processing (1/4)

Prior

Likelihood

$$p(\mathcal{D}|h) = \left[\frac{1}{size(h)}\right]^N = \left[\frac{1}{|h|}\right]^N,$$

$\mathcal{D}$ – observed data;
h – total number of items;
N – number of sampled (with replacement) items.

Occam's razor:
The model favors the simplest (smallest) hypothesis consistent with the data.

Jeffreys-Linley paradox:
Bayes-oriented decision systems will always favor the simpler model, since the probability of the observed data under a complex model with a very diffuse prior will be very small.

Posterior

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)p(h)}{\sum_{\acute{h}\in\mathcal{H}} p(D,\acute{h})} = \frac{p(h)\mathbb{I}(\mathcal{D}\in h)/|h|^N}{\sum_{\acute{h}\in\mathcal{H}} p(\acute{h})\mathbb{I}(\mathcal{D}\in \acute{h})/|\acute{h}|^N}$$

where $\mathbb{I}(\mathcal{D}\in h)$is 1 ***if and only if*** all the data are in extension of the hypothesis h.

Prior and posterior processing (2/4)

If we have enough data, the posterior $p(h|\mathcal{D})$ becomes ***Maximum A Posterior Estimate*** (***MAP***):

$$p(\mathcal{D}|h) \to \delta_{\hat{h}^{MAP}}(h),$$

$\hat{h}^{MAP} = \text{argmax}_h p(h|\mathcal{D})$ – the posterior mode;

$\delta_x(A) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$ – Dirac measure.

Note that the MAP estimate can be written as:

$$\hat{h}^{MAP} = \underset{h}{\text{argmax}}[p(\mathcal{D}|h)p(h)] = \underset{h}{\text{argmax}}\left[\log\left(p(\mathcal{D}|h)\right) + \log\left(p(h)\right)\right].$$

As we get more and more data, the MAP estimate converges towards the ***Maximum Likelihood Estimate*** (***MLE***):

$$\hat{h}^{MLE} = \underset{h}{\text{argmax}}[p(\mathcal{D}|h)] = \underset{h}{\text{argmax}}\left[\log\left(p(\mathcal{D}|h)\right)\right].$$

Prior and posterior processing (3/4)

The way to test if our beliefs are justified is to use them to predict objectively observable quantities with usage of ***posterior predictive distribution***:

$$p(\tilde{x} \in C|\mathcal{D}) = \sum_h p(y = 1|\tilde{x}, h)p(h|\mathcal{D}) .$$

When we have a small and/or ambiguous dataset, the posterior $p(h|\mathcal{D})$ is vague, which induces a broad predictive distribution. However, once we have "figured things out", the posterior becomes a delta function centered at the MAP estimate. In this case, we can use ***plug-in approximation***:

$$p(\tilde{x} \in C|\mathcal{D}) = \sum_h p(\tilde{x}|\, h)\delta_{\hat{h}}(h) = p\left(\tilde{x}|\hat{h}\right) .$$

Prior and posterior processing (4/4)

In general, computing the Maximum a Posterior Estimate $p(\mathcal{D}|h)$ can be quite difficult. One simple but popular approximation is known as the ***Bayesian information criterion*** (***BIC***)

$$BIC \triangleq \log p\left(\mathcal{D}|\widehat{\boldsymbol{\theta}}\right) - \frac{dof\left(\widehat{\boldsymbol{\theta}}\right)}{2}\log N \approx \log p(\mathcal{D})\,,$$

where $\widehat{\boldsymbol{\theta}}$ – maximum likelihood estimation of used model parameters; $dof\left(\widehat{\boldsymbol{\theta}}\right)$ – the number of degrees of freedom in used model.

The BIC method is very closely related to the ***Minimum Description Length*** or ***MDL*** principle, which characterizes the score of a model in terms of how well it fits the data, minus how complex the model is to define.

A very similar expression of BIC / MDL is called the ***Akaike information criterion*** or ***AIC***

$$AIC(m,\mathcal{D}) \triangleq \log p\left(\mathcal{D}|\widehat{\boldsymbol{\theta}}_{MLE}\right) - dof(m)\,.$$

Examples. Beta-binomial model (1/3)

Suppose $X_i \sim Ber(\theta)$, where $X_i = 1$ represents "heads", $X_i = 0$ represents "tails", and $\theta \in [0; 1]$ is the rate parameter (probability of head). If the data are iid, the likelihood has the form:

$$p(\mathcal{D}|\theta) = \theta^{N_1}(1-\theta)^{N_0}.$$

where we have $N_1 = \sum_{i=1}^{N} \mathbb{I}(x_i = 1)$ heads and $N_0 = \sum_{i=1}^{N} \mathbb{I}(x_i = 0)$ tails, $N = N_0 + N_1$ is observed trials. In this case we have $N_1 \sim Bin(N, \theta)$, which has following pdf:

$$Bin(k|n, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k}.$$

Since binomial coefficients $\binom{n}{k}$ is a constant independent of θ, the ***likelihood of the binomial sampling model*** is the same as the likelihood for the Bernoulli model – any inference we have about θ will be same whether we observe the counts $\mathcal{D} = (N_1, N)$ or sequence of trials $\mathcal{D} = \{x_1, \cdots x_N\}$.

Examples. Beta-binomial model (2/3)

To make the math easier, it would be convenient if the prior had the same form as the likelihood:

$$p(\theta) \propto \theta^{\gamma_1}(1-\theta)^{\gamma_0}$$

for some prior parameters γ_1 and γ_2. Then we could easily evaluate the posterior by simply adding the exponents

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta) = \theta^{N_1}(1-\theta)^{N_0}\theta^{\gamma_1}(1-\theta)^{\gamma_0} = \theta^{N_1+\gamma_1}(1-\theta)^{N_0+\gamma_0}$$

When the prior and the posterior have the same form, we say that the prior is a ***conjugate prior*** for the corresponding likelihood. In case of the Bernoulli, the conjugate prior is the beta distribution:

$$Beta(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Then posterior is

$$p(\theta|\mathcal{D}) \propto Bin(N_1|\theta, N_0+N_1)Beta(\theta|a,b) \propto Beta(\theta|N_1+a, N_0+b)$$

Examples. Beta-binomial model (3/3)

Consider predicting the probability of heads in a single future trial under a $\mathrm{Beta}(a, b)$ posterior:

$$p(\tilde{x} = 1|\mathcal{D}) = \int_0^1 p(x = 1|\theta)p(\theta|\mathcal{D})d\theta = \int_0^1 \theta \mathrm{Beta}(\theta|a, b)d\theta = \mathbb{E}[\theta|\mathcal{D}] = \frac{a}{a+b}.$$

Zero count (sparse data) problem – occurs when estimating counts from small amount of data;

Black swan paradox – problem of how to draw general conclusion about the future from specific observation from the past.

Suppose now we were interested in predicting the number of heads, x, in M future trials:

$$p(x|\mathcal{D}, M) = \int_0^1 \mathrm{Bin}(x|\theta, M)\mathrm{Beta}(\theta|a, b)d\theta =$$

$$\binom{M}{x}\frac{1}{B(a,b)}\int_0^1 \theta^x(1-\theta)^{M-x}\theta^{a-1}(1-\theta)^{b-1}d\theta = \binom{M}{x}\frac{B(x+a, M-x+b)}{B(a,b)}.$$

Naïve Bayes classifier

Let us classify vector of discrete-valued features $\mathbf{x} \in \{1, \cdots, K\}^D$, where K is number of values for each feature, D is the number of features. The simplest approach is to assume the features are conditionally independent given the class label:

$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^{D} p(x_j|y = c, \boldsymbol{\theta}_{jc}).$$

The model is called "naïve" since we do not expect the features to be independent, even conditional on the class label. One reason of successful application of naïve Bayes classifier is that the model is quite simple, and hence it is relatively immune to overfitting.

Type of feature	Recommended type of class-conditional density
Real-valued	$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^{D} \mathcal{N}(x_j|\mu_{jc}, {\sigma_{jc}}^2)$
Binary	$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^{D} \mathrm{Ber}(x_j|\mu_{jc})$
Categorical	$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^{D} \mathrm{Multinoulli}(x_j|\boldsymbol{\mu}_{\mathbf{jc}})$

Bayesian approach applications

Hierarchical Bayes (multi-level model) based on putting a prior on used prior:

$$\eta \rightarrow \theta \rightarrow \mathcal{D}.$$

Empirical Bayes violates the principle that the prior should be chosen independently of the data:

$$\hat{\eta} = \operatorname{argmax} p(\mathcal{D}|\boldsymbol{\eta}) = \operatorname{argmax} \left[\int p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\eta}) d\boldsymbol{\theta} \right].$$

Method	Definition
Maximum likelihood	$\widehat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})$
MAP estimation	$\widehat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\eta})$
Empirical Bayes	$\widehat{\boldsymbol{\eta}} = \operatorname{argmax}_{\boldsymbol{\eta}} p(\mathcal{D}|\boldsymbol{\eta})$
Full Bayes	$p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\eta}) p(\boldsymbol{\eta})$

Conclusion

- Concept of Bayesian learning was considered;
- Methods for processing the prior and posterior information were presented;
- Practical applications of Bayesian approach were shown.