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

Lecture 4

Generative Models for Real-World Systems. Frequentist Concept

Dmytro Progonov,

PhD, Associate Professor,

Department Of Physics and Information Security Systems

Content

- Frequentist statistics concept;
- Estimator risk minimization procedures;
- Estimator’s properties;
- Pathologies of frequentist statistics.

Frequentist statistics concept

Frequentist (***classical*** or ***orthodox***) ***statistics*** are based on the concept of a ***sampling distribution***. This is the distribution that an ***estimator*** δ has when applied to the multiple dataset $\{\mathcal{D}_1, \mathcal{D}_2 \cdots \mathcal{D}_K\}$ sampled from the true but unknown distributions.

Parameter estimate – $\hat{\theta} = \delta(\mathcal{D})$.

In classical (frequentist) decision theory, there is a loss function and a likelihood, but there is no prior and hence no posterior or posterior expected loss. Thus *there is no automatic way to deriving an optimal estimator, unlike the Bayesian case*.

Having chosen an estimator δ, we define its expected ***loss*** or ***risk*** as

$$\boxed{R(\theta^*, \delta) \triangleq \mathbb{E}_{p(\widetilde{\mathcal{D}}|\theta^*)}\left[L\left(\theta^*, \delta\left(\widetilde{\mathcal{D}}\right)\right)\right] = \int L\left(\theta^*, \delta\left(\widetilde{\mathcal{D}}\right)\right) p\left(\widetilde{\mathcal{D}}|\theta^*\right) d\,\widetilde{\mathcal{D}},}$$

where $\widetilde{\mathcal{D}}$ is a data sampled from "nature's distribution", which is represented by parameter θ^*.

Estimator risk minimization procedures. Bayes risk

One approach to convert risk $R(\theta^*, \delta)$ into a single measure of quantity, $R(\delta)$, which does not depend on knowing θ^*, is to put a prior on θ^*, and then to define ***Bayes (integrated) risk***:

$$\boxed{R_B(\delta) \triangleq \mathbb{E}_{p(\theta^*)}[R(\theta^*, \delta)] = \int R(\theta^*, \delta) p(\theta^*) d\theta^*}$$

A ***Bayes estimator*** or ***Bayes decision rule*** is one which minimizes the expected risk:

$$\boxed{\delta_B \triangleq \underset{\delta}{\operatorname{argmin}}\, R_B(\delta)}$$

Theorem: A Bayes estimator can be obtained by minimizing the posterior expected loss for each $\mathbf{x}$.

Theorem (Wald, 1950): Every admissable decision rule is a Bayes decision rule with resoect to some, possibly improper, prior distribution.

Estimator risk minimization procedures. Minimax risk

An alternative approach to Bayes estimator, that does not require choice of prior, is usage of minimax rule:

$$\boxed{\delta_{MM} \triangleq \underset{\delta}{\operatorname{argmin}}\, R_{max}(\delta),}$$

where $R_{max}(\delta) \triangleq \max_{\delta} R(\theta^*, \delta)$ – the maximum risk of an estimator.

Minimax estimator is very pessimistic – one can show that all minimax estimators are equivalent to Bayes estimators under a ***least favorable prior***.

Estimator risk minimization procedures. Empirical risk minimization

We define the empirical risk as follows:

$$\boxed{R_{emp}(\mathcal{D},\delta) \triangleq R\big(p_{emp}(\cdot\,|\mathcal{D}),\delta\big) = \frac{1}{N}\sum_{i=1}^{N} L(y_i,\delta(\mathbf{x}_i))}$$

A ***empirical risk minimization*** is task of finding a decision procedure to minimize the empirical risk:

$$\boxed{\delta_{ERM} \triangleq \underset{\delta}{\operatorname{argmin}}\, R_{emp}(\mathcal{D},\delta)}$$

Possible approaches for solving the mentioned task:

Regularized risk minimization:

$$\acute{R}(\mathcal{D},\delta) = R_{emp}(\mathcal{D},\delta) + \lambda C(\delta)$$

Structural risk minimization:

$$\begin{cases} \acute{\delta}_\lambda = \underset{\delta}{\operatorname{argmin}}\big[R_{emp}(\mathcal{D},\delta) + \lambda C(\delta)\big] \\ \acute{\lambda} = \underset{\lambda}{\operatorname{argmin}}\, \acute{R}\big(\acute{\delta}_\lambda\big) \end{cases}$$

Estimator’s properties

Consistent estimator – if it eventually recovers the true parameters that generated the data as the sample size goes to infinity, i.e. $\hat{\theta}(\mathcal{D}) \to \theta^*$ as $|\mathcal{D}| \to \infty$ (convergence in probability sense).

Unbiased estimator provides the null bias of parameters measurement:

$$bias\left(\hat{\theta}(\cdot)\right) = \mathbb{E}_{p(\mathcal{D}|\theta_*)}\left[\hat{\theta}(\mathcal{D}) - \theta_*\right] \to 0$$

Minimum variance estimator achieves the lower bound on the variance of any unbiased estimator:

Cramer-Rao inequality: Let $X_1, X_2 \cdots X_n \sim p(X|\theta_0)$ and $\hat{\theta} = \hat{\theta}(x_1, x_2 \cdots x_n)$ be an unbiased estimator of θ_0. Then, under various smoothness assumption on $p(X|\theta_0)$ we have

$$var\left[\hat{\theta}\right] \geq \frac{1}{nI(\theta_0)},$$

where $I(\theta_0)$ is Fisher information matrix.

$$\boxed{MSE = variance + bias^2}$$

Pathologies of frequentist statistics. Counter-intuitive behavior of confidence intervals

A frequentist confidence interval for some parameter θ is defined y expression:

$$\boxed{C_\alpha(\theta) = (l, u): P\left(l(\tilde{\mathcal{D}}) \le \theta \le u(\tilde{\mathcal{D}}) | \tilde{\mathcal{D}} \sim \theta\right) = 1 - \alpha.}$$

In Bayesian statistics, we condition on what is known – namely the observed data, $\mathcal{D}$ – and average on what is unknown, namely the parameter θ.

In frequentist statistics, we do exactly the opposite: we condition on what is unknown – namely the true parameter value θ – and average over hypothetical future dataset $\tilde{\mathcal{D}}$.

Pathologies of frequentist statistics. P-values considered harmful

Suppose we want to decide whether to accept/reject some baseline model (*null hypothesis*). In frequentist statistics, it is standard to compute the ***p-value***, which is defined as probability (under the null) of observing some test statistics $f(\mathcal{D})$:

$$\boxed{pvalue(\mathcal{D}) \triangleq P\left(f\left(\widetilde{\mathcal{D}}\right) \geq f(\mathcal{D}) | \widetilde{\mathcal{D}} \sim H_0\right)}$$

Given the p-value, we reject the null hypothesis iff the p-value is less than some threshold (such as $\alpha = 0.05$). If we reject it, we say the difference between observed and expected test statistics is ***statistically significant*** at level α.

This procedure guarantees that our expected type I (false positive) error rate is at most α. This sometimes is interpreted as saying that *frequentist hypothesis testing is very conservative* (unlikely to accidently reject the null hypothesis). In fact the opposite is the case – because this method only worries about trying to reject the null, it can never gather evidence in favor of the null, no matter how large the sample size.

Another problem with *p-values* is that their *computation depends on decision you make about when to stop collecting data*, even if these decisions do not change the data you actually observed.

Pathologies of frequentist statistics.
The likelihood principle

The fundamental reason for mentioned pathologies is that frequentist statistics violates:

Likelihood principle – interference should be based on likelihood of the observed data, not based on hypothetical future data that you have not observed.

Sufficiency principle – sufficient statistics contains all the relevant information about an unknown parameter.

Weak conditionality – interference should be based on events that happened, not which might have happened.

Conclusion

- Concept of frequentist statistics was presented;
- Estimators and theirs properties were considered;
- Procedures for estimator risk minimization were shown;
- Limitations of practical usage the frequentist statistics were considered.