Statistical Machine Learning (BE4M33SSU)
Lecture 8: Bayesian inference and learning

Czech Technical University in Prague

- Bayesian parameter estimation
- Mixtures of classifiers

When ERM and MLE fail

Empirical risk minimisation:

- The best attainable (Bayes) risk is $R^* = \inf_{h \in \mathcal{Y}^\mathcal{X}} R(h)$
- The best predictor in $\mathcal{H}$ is $h_\mathcal{H} \in \arg\min_{h \in \mathcal{H}} R(h)$
- The predictor h_m learned from $\mathcal{T}^m$ has risk $R(h_m)$

$$\underbrace{\Big(R(h_m) - R^*\Big)}_{\text{excess error}} = \underbrace{\Big(R(h_m) - R(h_\mathcal{H})\Big)}_{\text{estimation error}} + \underbrace{\Big(R(h_\mathcal{H}) - R^*\Big)}_{\text{approximation error}}$$

- Misspecified hypothesis space $\mathcal{H} \Rightarrow$ high approximation error
- Size of $\mathcal{T}^m$ too small $\Rightarrow$ high estimation error

Maximum likelihood estimate: similar

- Misspecified model class $p_\theta(x, y)$, $\theta \in \Theta$
- Size of $\mathcal{T}^m$ too small

Bayesian parameter estimation

Model class $p_\theta(x, y)$, $\theta \in \Theta$

- Interpret the unknown parameter $\theta \in \Theta$ as a random variable
- assume a prior distribution $p(\theta)$ for θ
- choose a loss incurred by wrong estimation, e.g. $\ell(\theta, \theta) = [\theta - \theta']^2$

The estimation is based on the posterior distribution for the parameter, i.e.

$$p(\theta \mid \mathcal{T}^m) = \frac{p(\mathcal{T}^m \mid \theta) p(\theta)}{p(\mathcal{T}^m)} = \frac{p(\mathcal{T}^m \mid \theta) p(\theta)}{\int_\Theta p(\mathcal{T}^m \mid \theta')\, p(\theta')\, d\theta'}$$

Bayes estimator

$$e_B(\mathcal{T}^m) = \operatorname*{arg\,min}_{\theta' \in \Theta} \int_\Theta p(\theta \mid \mathcal{T}^m) [\theta - \theta']^2 \, d\theta$$

For the considered squared-error loss we obtain

$$e_B(\mathcal{T}^m) = \theta^*(\mathcal{T}^m) = \int_\Theta p(\mathcal{T}^m \mid \theta) p(\theta)\, \theta \, d\theta$$

Bayesian parameter estimation

Notice how the posterior distribution

$$p(\theta \mid \mathcal{T}^m) \propto p(\mathcal{T}^m \mid \theta)\, p(\theta)$$

interpolates between the situation without any training data, i.e. $m = 0$ and the log-likelihood of training data for $m \to \infty$.

Bayesian risk minimisation

Is it possible to consider a similar approach for hypothesis learning? Yes.

- Model class $p(x, y \mid \theta)$, $\theta \in \Theta$
- Prior distribution $p(\theta)$ on Θ
- Prediction strategy $h \colon \mathcal{X} \to \mathcal{Y}$
- A loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Given training data $\mathcal{T}^m = \big\{(x^i, y^i) \,\big|\, i = 1, \ldots, m\big\}$ compute the posterior probability to observe a pair (x, y) by marginalising over $\theta \in \Theta$:

$$p(x, y \mid \mathcal{T}^m) = \frac{1}{p(\mathcal{T}^m)} \int_{\Theta} p(\mathcal{T}^m \mid \theta)\, p(x, y \mid \theta)\, p(\theta)\, d\theta$$

Define the Bayes risk of a strategy h by

$$R(h, \mathcal{T}^m) \propto \sum_{x,y} \int_{\Theta} p(\mathcal{T}^m \mid \theta)\, p(x, y \mid \theta)\, p(\theta)\, \ell(y, h(x))\, d\theta$$

Bayesian risk minimisation

For 0-1 loss this leads to the predictor

$$h(x, \mathcal{T}^m) = \arg\max_{y \in \mathcal{Y}} \int_{\Theta} \underbrace{p(\theta)\, p(\mathcal{T}^m \mid \theta)}_{\alpha(\theta)}\, p(x, y \mid \theta)\, d\theta = \arg\max_{y \in \mathcal{Y}} \int_{\Theta} \alpha(\theta) p(x, y \mid \theta)\, d\theta$$

which means to find the optimal predictor for a model mixture.