Statistical Machine Learning

Lecture 2: More Advanced Model Selection

2022-23

Introduction

- The solution to the regression problem is the estimation of the underlying generator of data.
- The most general description of the data generator is given in terms of the probability density $\mathcal{P}(t, \boldsymbol{x})$, where t is the dependent variable (the output of the network) and $\boldsymbol{x}$is the independent variable (input or feature).
- By definition we have:

$$\mathcal{P}(t, \boldsymbol{x}) = \mathcal{P}(t|\boldsymbol{x})\mathcal{P}(\boldsymbol{x})$$
$$\mathcal{P}(\boldsymbol{x}) = \int \mathrm{d}t\mathcal{P}(t, \boldsymbol{x}).$$

- In order to make a prediction (on an output given an input) we need to model $\mathcal{P}(t|\boldsymbol{x})$.

Likelihood

- Several error measures are based on the *likelihood* $\mathcal{L}(\mathcal{D})$ of the data set $\mathcal{D} = \{(t_n, \boldsymbol{x}_n)\}_{n=1}^N$:

$$\mathcal{L}(\mathcal{D}) = \prod_n \mathcal{P}(t_n, \boldsymbol{x}_n)$$

  where we have assumed that the data points are drawn independently from the same distribution.
- Maximizing the likelihood is equivalent to minimizing the error (or energy) defined as:

$$E = -\ln \mathcal{L} = -\sum_n \ln \mathcal{P}(t_n|\boldsymbol{x}_n) - \sum_n \ln \mathcal{P}(\boldsymbol{x}_n).$$

- The second term to the right hand side does not depend on the machine learning model being used, thus

$$E' = -\sum_n \ln \mathcal{P}(t_n|\boldsymbol{x}_n), \tag{1}$$

Gaussian Noise

- Suppose the variable t is given by a combination of a deterministic process $h(\boldsymbol{x})$ plus a random variable ϵ drawn from a Gaussian distribution with zero mean and variance σ^2:

$$t = h(\boldsymbol{x}) + \epsilon$$

$$\mathcal{P}(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$$

Gaussian Noise

- The deterministic function $h(\boldsymbol{x})$ is unknown, but is the only contribution to t that can be inferred from the data. Let as assume that there is an estimate $f(\boldsymbol{x};\boldsymbol{w})$ that implements a model for $h(\boldsymbol{x})$ (one estimate we have explored is the least-square polynomial, the vector $\boldsymbol{w}$ represents the parameter of the polynomial). Such a model is associated with the following conditional probability of t :

$$\mathcal{P}(t|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{[t - f(\boldsymbol{x};\boldsymbol{w})]^2}{2\sigma^2}\right\}. \tag{2}$$

Gaussian Noise and Maximum Likelihood

- By applying (1) with (2) we have that the log likelihood for a model with Gaussian Noise gives:

$$E' = \frac{1}{2\sigma^2} \sum_n [t_n - f(\boldsymbol{x}_n; \boldsymbol{w})]^2 + \frac{N}{2} \ln(2\pi\sigma^2)$$

- The first term of the right hand side is the usual sum-of-squares error.
- Once optimized the model, by solving $\nabla_{\boldsymbol{w}} E = \mathbf{0}$, we can demonstrate that the variance satisfies:

$$\sigma^2 = \frac{1}{N} \sum_n [t_n - f(\boldsymbol{x}_n; \boldsymbol{w}^\star)]^2$$

where $\boldsymbol{w}^\star$ is the solution of $\nabla_{\boldsymbol{w}} E = \mathbf{0}$.

Noisy data

- We consider the cost function to be the sum of squares and that the size of the data set is *large*:

$$E(\boldsymbol{w}) = \lim_{N\to\infty} \frac{1}{2N} \sum_{n=1}^{N} [f(\boldsymbol{x}_n; \boldsymbol{w}) - t_n]^2 ,$$

where $f(\bullet; \boldsymbol{w}) : \mathbb{R}^d \to \mathbb{R}$ is the function implemented by the network with parameters $\boldsymbol{w} \in \mathbb{R}^d$.

- In such a limit we have that:

$$E(\boldsymbol{w}) = \frac{1}{2} \int \mathrm{d}t \, \mathrm{d}\boldsymbol{x} \, \mathcal{P}(t|\boldsymbol{x})\mathcal{P}(\boldsymbol{x}) \, [f(\boldsymbol{x}; \boldsymbol{w}) - t]^2$$

- Let us define the conditional averages:

$$\mathbb{E}[t|\boldsymbol{x}] \equiv \int \mathrm{d}t \, \mathcal{P}(t|\boldsymbol{x}) y, \qquad \mathbb{E}[t^2|\boldsymbol{x}] \equiv \int \mathrm{d}t \, \mathcal{P}(t|\boldsymbol{x}) t^2$$

Noisy data

- Then

$$
\begin{aligned}
E(\boldsymbol{w}) = \frac{1}{2} \int \mathrm{d}t\, \mathrm{d}\boldsymbol{x}\, \mathcal{P}(t|\boldsymbol{x})\mathcal{P}(\boldsymbol{x}) \Big\{ [f(\boldsymbol{x};\boldsymbol{w}) - \mathbb{E}[t|\boldsymbol{x}]]^2 \\
+ 2\,[f(\boldsymbol{x};\boldsymbol{w}) - \mathbb{E}[t|\boldsymbol{x}]]\, [\mathbb{E}[t|\boldsymbol{x}] - t] + [\mathbb{E}[t|\boldsymbol{x}] - t]^2 \Big\} \\
= \frac{1}{2} \int \mathrm{d}\boldsymbol{x}\, \mathcal{P}(\boldsymbol{x})\, [f(\boldsymbol{x};\boldsymbol{w}) - \mathbb{E}[t|\boldsymbol{x}]]^2 + \qquad (3) \\
+ \frac{1}{2} \int \mathrm{d}\boldsymbol{x}\, \mathcal{P}(\boldsymbol{x}) \left[\mathbb{E}[t^2|\boldsymbol{x}] - \mathbb{E}[t|\boldsymbol{x}]^2 \right]. \qquad (4)
\end{aligned}
$$

- Observe that the second contribution (4) is positive and does not depend on the parameters $\boldsymbol{w}$.
- The minimization of E is achieved for $\boldsymbol{w}^\star \in \mathbb{R}^d$ such that $f(\boldsymbol{x};\boldsymbol{w}^\star) = \mathbb{E}[t|\boldsymbol{x}]$.

Finite data set

- Suppose that $|\mathcal{D}| = N < \infty$. In such a case, the quantity $[f(\boldsymbol{x};\boldsymbol{w}) - \mathbb{E}[t|\boldsymbol{x}]]^2$ depends on the particular data set $\mathcal{D}$ used to train the model.
- We can eliminate this dependency by averaging over all possible data sets $\mathcal{D}$ with cardinality N. We denote such an average by $\mathbb{E}_{\mathcal{D}}[\cdot]$.
- Then:

$$
\begin{aligned}
(f(\boldsymbol{x};\boldsymbol{w}) - \mathbb{E}[t|\boldsymbol{x}])^2 &= (f(\boldsymbol{x};\boldsymbol{w}) - \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\boldsymbol{w})] + \\
&+ \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\boldsymbol{w})] - \mathbb{E}[t|\boldsymbol{x}])^2 \\
&= (f(\boldsymbol{x};\boldsymbol{w}) - \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\boldsymbol{w})])^2 + \\
&+ (\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\boldsymbol{w})] - \mathbb{E}[t|\boldsymbol{x}])^2 \\
&+ 2\,(f(\boldsymbol{x};\boldsymbol{w}) - \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\boldsymbol{w})]) \times \\
&\times (\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\boldsymbol{w})] - \mathbb{E}[t|\boldsymbol{x}])
\end{aligned}
$$

Finite data set

- By averaging both member over $\mathcal{D}$:

$$\mathbb{E}_{\mathcal{D}}\left[[f(\boldsymbol{x};\boldsymbol{w}) - \mathbb{E}[t|\boldsymbol{x}]]^2\right] = (\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\boldsymbol{w})] - \mathbb{E}[t|\boldsymbol{x}])^2 + \quad (5)$$

$$+ \mathbb{E}_{\mathcal{D}}\left[(f(\boldsymbol{x};\boldsymbol{w}) - \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\boldsymbol{w})])^2\right], \quad (6)$$

  where (5) is the squared *bias* term and (6) the *variance* term.
- The bias measures the extent to which the average over all data sets $\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\boldsymbol{w})]$ differs from the desired function $\mathbb{E}[t|\boldsymbol{x}]$.
- The variance measures the extent to which the network function $f(\boldsymbol{x};\boldsymbol{w})$ is sensitive to the particular choice of data set.
- Both contributions depend on $\boldsymbol{x}$.

Bias vs Variance

- We can eliminate the dependency over $\boldsymbol{x}$ by integrating:

$$\begin{aligned}
(\mathrm{bias})^2 &= \frac{1}{2}\int \mathrm{d}\boldsymbol{x}\mathcal{P}(\boldsymbol{x})\left(\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\boldsymbol{w})] - \mathbb{E}[t|\boldsymbol{x}]\right)^2 \\
\mathrm{variance} &= \frac{1}{2}\int \mathrm{d}\boldsymbol{x}\mathcal{P}(\boldsymbol{x})\mathbb{E}_{\mathcal{D}}\left[(f(\boldsymbol{x};\boldsymbol{w}) - \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\boldsymbol{w})])^2\right].
\end{aligned}$$

- Increasing the complexity of the model (number of parameters) reduces the bias but increase the sensibility of the model (variance) to the data set used (over fitting).

Information Criteria

- Let us define the Kullback-Liebler (KL) divergence as the functional $I : \mathbb{D} \times \mathbb{D} \to \mathbb{R}^+ \cup \{0\}$, where $\mathbb{D}$ is the space of functions that are positive and integrable (i.e. suitable probability distributions), as:

$$I[f, g] = \int \mathrm{d}x\, f(x) \ln \left(\frac{f(x)}{g(x)} \right).$$

- The KL divergency is positive: By using that $\ln a \leq a - 1$ for all $a > 0$

$$\int \mathrm{d}x\, f(x) \ln \left(\frac{f(x)}{g(x)} \right) \geq \int \mathrm{d}x\, f(x) \left(1 - \frac{g(x)}{f(x)} \right) = 0.$$

- Suppose f is the distribution of the data (inaccessible) and g is the model you are using to estimate f. $I[f, \cdot]$ can be used to compare different models g_i and choose which one is the *closest* to f.

AIC and BIC Scores

- The Akaike Information Criterion is an estimate of the KL divergency

$$AIC = 2K - 2\ln[\mathcal{L}(\boldsymbol{w}^\star; \mathcal{D}_n)]$$

  where K is the number of parameters used in tho model and $\boldsymbol{w}^\star$ is the estimate of the parameter $\boldsymbol{w}$ that maximizes the likelihood.
- The Bayesian Information Criterion is an improved (more sensitive) version of the AIC:

$$BIC = \ln(n)K - 2\ln[\mathcal{L}(\boldsymbol{w}^\star; \mathcal{D}_n)]$$

  where n is the number of data points.
- In both cases we have a score based on the balance between the model complexity (the first term) and the model performance.

Maximum Likelihood Revisited

- By the Bayes Theorem we have that

$$p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M})p(\mathcal{M})}{p(\mathcal{D})}$$

  where $\mathcal{M}$represents a given model or process, $\mathcal{D}$ is the data set, or observations, $p(\mathcal{M})$ is the density of probability of the model, before we have access to the data (known as prior), $p(\mathcal{D}|\mathcal{M})$ is the conditional probability of the data given the model. But for a fixed set of data, this is the likelihood of the model given the data. $p(\mathcal{D})$ is the marginal probability of the data that in this scheme plays the role of a normalization constant. $p(\mathcal{M}|\mathcal{D})$ is the probability of the model given the data. This is known as the posterior and represents an update of the prior $p(\mathcal{M})$ after the data $\mathcal{M}$ is aquired.

- If the model $\mathcal{M}$ depends itself on parameters $\boldsymbol{w}$ that are also distributed variables (drawn from a distribution $g(\cdot)$) we can write:

$$p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{M}) \int \mathrm{d}\boldsymbol{w} g(\boldsymbol{w}) p(\mathcal{D}|\mathcal{M}, \boldsymbol{w})}{p(\mathcal{D})}.$$

Partial Demonstration of the BIC score

- Given the data set $\mathcal{D}_n = \{\boldsymbol{x}_j\}_{j=1}^n$ composed by n (large) independent and identically-distributed (iid) observations $\boldsymbol{x}_j \in \mathbb{R}^d$, and a model characterized by a density distribution $p(\boldsymbol{x}|\boldsymbol{w})$, where $\boldsymbol{w} \in \mathbb{R}^K$ is the set of parameters used by the model, the -log-likelihood is given by

$$-\ln p(\mathcal{D}_n|\mathcal{M}, \boldsymbol{w}) = -\sum_{j=1}^{n} \ln p(\boldsymbol{x}_j|\boldsymbol{w}). \tag{7}$$

- We assume that there exists a vector $\boldsymbol{w}^\star$ such that expression (7) is minimized:

$$\boldsymbol{w}^\star = \operatorname{argmin}_{\boldsymbol{w}\in\mathbb{R}^K} \left(-\ln p(\mathcal{D}_n|\mathcal{M}, \boldsymbol{w})\right).$$

Partial Demonstration of the BIC score

- We also assume that for sufficiently large number of observations n the meaningful parameters will be concentrated close to $\boldsymbol{w}^\star$, which justifies athe Taylor expansion:

$$-\ln p(\mathcal{D}_n|\mathcal{M}, \boldsymbol{w}) = -\ln p(\mathcal{D}_n|\mathcal{M}, \boldsymbol{w}^\star) + \frac{1}{2}\delta\boldsymbol{w}^T \boldsymbol{I}_n \delta\boldsymbol{w},$$

where $\delta\boldsymbol{w} = \boldsymbol{w} - \boldsymbol{w}^\star$ and

$$[\boldsymbol{I}_n]_{\ell,k} = \frac{\partial^2}{\partial w_\ell \partial w_k}\left[-\sum_{j=1}^{n} \ln p(\boldsymbol{x}_j|\boldsymbol{w})\right]\Bigg|_{\boldsymbol{w}=\boldsymbol{w}^\star}$$

is the matrix of second derivatives (Hessian). This matrix is positive definite therefore its eigenvalues are positive.

Partial Demonstration of the BIC score

- By the law of large numbers, for sufficiently large n, we have that

$$[\boldsymbol{I}_n]_{\ell,k} \to -n \frac{\partial^2}{\partial w_\ell \partial w_k} \ln p(\boldsymbol{x}|\boldsymbol{w}^\star) = n[\boldsymbol{I}]_{\ell,k}$$

- By Bayes we have that, for sufficiently large n,

$$\begin{aligned}
-\ln p(\mathcal{M}|\mathcal{D}_{\boldsymbol{n}}) &= \ln p(\mathcal{D}_{\boldsymbol{n}}) - \ln p(\mathcal{M}) - \ln \int \mathrm{d}\boldsymbol{w} g(\boldsymbol{w}) p(\mathcal{D}_{\boldsymbol{n}}|\mathcal{M}, \boldsymbol{w}^\star) \exp\left(-\frac{\boldsymbol{n}}{2} \delta \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{I} \delta \boldsymbol{w}\right) \\
&= \ln p(\mathcal{D}_{\boldsymbol{n}}) - \ln p(\mathcal{M}) - \ln \left\{ p(\mathcal{D}_{\boldsymbol{n}}|\mathcal{M}, \boldsymbol{w}^\star) \sqrt{\frac{(2\pi)^{\boldsymbol{K}}}{\boldsymbol{n}^{\boldsymbol{K}} \det(\boldsymbol{I})}} \int \mathrm{d}\boldsymbol{w} g(\boldsymbol{w}) \mathcal{N}(\boldsymbol{w}|\boldsymbol{w}^\star, \boldsymbol{I}^{-1}/\boldsymbol{n}) \right\} \\
&= \ln p(\mathcal{D}_{\boldsymbol{n}}) - \ln p(\mathcal{M}) - \ln \left\{ \frac{p(\mathcal{D}_{\boldsymbol{n}}|\mathcal{M}, \boldsymbol{w}^\star)}{\boldsymbol{n}^{\boldsymbol{K}/2}} g(\boldsymbol{w}^\star) \sqrt{\frac{(2\pi)^{\boldsymbol{K}}}{\det(\boldsymbol{I})}} \right\} \\
&= -\ln p(\mathcal{D}_{\boldsymbol{n}}|\mathcal{M}, \boldsymbol{w}^\star) + \frac{K}{2} \ln(n) - \ln \left\{ g(\boldsymbol{w}^\star) \sqrt{\frac{(2\pi)^{\boldsymbol{K}}}{\det(\boldsymbol{I})}} \frac{p(\mathcal{M})}{p(\mathcal{D}_{\boldsymbol{n}})} \right\}
\end{aligned}$$

Partial Demonstration of the BIC score

- Then

$$-\ln p(\mathcal{M}|\mathcal{D}_n) = -\ln p(\mathcal{D}_n|\mathcal{M}, \boldsymbol{w}^\star) + \frac{K}{2}\ln(n) + \mathrm{OT}$$

- Observe that the larger the number of parameters (K) the smaller the error ($-\ln p(\mathcal{D}_n|\mathcal{M}, \boldsymbol{w}^\star)$) but the larger the second term.