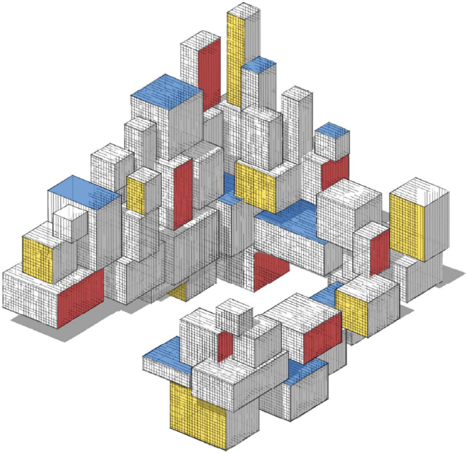
Unsupervised Learning



Purpose

In this lecture we discuss the statistical learning framework for unsupervised learning. Topics include:

- Cross-entropy risk
- Fisher information matrix
- Information matrix equality
- Akaike information criterion
- Fisher's scoring method

Unsupervised Learning

In unsupervised learning there is no “response” variable and the overall goal is to extract useful information and patterns from the data, e.g., in the form $\tau = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ or as a matrix $\mathbf{X}^\top = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$.

In essence, the objective of unsupervised learning is to learn about the underlying probability distribution of the data.

We set up a framework for unsupervised learning that is similar to the framework used for supervised learning, involving risk and loss minimization.

However, we use here the cross-entropy risk rather than the squared-error risk.

Risk and Loss in Unsupervised Learning

In unsupervised learning, the training data $\mathcal{T} := \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ only consists of (what are usually assumed to be) independent copies of a feature vector $\boldsymbol{X}$; there is no response data.

Suppose our objective is to learn the unknown pdf f of $\boldsymbol{X}$ based on an outcome $\tau = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of the training data $\mathcal{T}$.

Conveniently, we can follow the same line of reasoning as for *supervised* learning.

For some loss function Loss, we wish to find a pdf g that best approximates the pdf f in terms of minimizing a risk

$$\ell(g) := \mathbb{E}\,\mathrm{Loss}(f(\boldsymbol{X}), g(\boldsymbol{X}))$$

We already encountered the Kullback–Leibler risk

$$\ell(g) := \mathbb{E} \ln \frac{f(\boldsymbol{X})}{g(\boldsymbol{X})} = \mathbb{E} \ln f(\boldsymbol{X}) - \mathbb{E} \ln g(\boldsymbol{X}).$$

Cross-Entropy Risk

If $\mathcal{G}$ is a class of functions that contains f, then minimizing the Kullback–Leibler risk over $\mathcal{G}$ will yield the (correct) minimizer f.

Since the term $\mathbb{E} \ln f(\boldsymbol{X})$ does not depend on g, it plays no role in the minimization of the Kullback–Leibler risk.

By removing this term, we obtain the cross-entropy risk (for discrete $\boldsymbol{X}$ replace the integral with a sum):

$$\ell(g) := -\mathbb{E} \ln g(\boldsymbol{X}) = -\int f(\boldsymbol{x}) \ \ln g(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

Thus, minimizing the cross-entropy risk over all $g \in \mathcal{G}$, again gives the minimizer f, provided that $f \in \mathcal{G}$.

Cross-Entropy Training Loss

Unfortunately, solving this problem is also infeasible in general, as it still depends on f.

Instead, we seek to minimize the cross-entropy training loss:

$$\ell_\tau(g) := \frac{1}{n}\sum_{i=1}^{n} \text{Loss}(f(\boldsymbol{x}_i), g(\boldsymbol{x}_i)) = -\frac{1}{n}\sum_{i=1}^{n} \ln g(\boldsymbol{x}_i)$$

for functions $g \in \mathcal{G}$, where $\tau = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is an iid sample from f.

This optimization problem is equivalent to solving the maximization problem

$$\max_{g \in \mathcal{G}} \sum_{i=1}^{n} \ln g(\boldsymbol{x}_i).$$

Table: Summary of definitions for unsupervised learning.

$\boldsymbol{x}$	Fixed feature vector.
$\boldsymbol{X}$	Random feature vector.
$f(\boldsymbol{x})$	Pdf of $\boldsymbol{X}$ evaluated at the point $\boldsymbol{x}$.
τ or τ_n	Fixed training data $\{\boldsymbol{x}_i, i = 1, \ldots, n\}$.
$\mathcal{T}$ or $\mathcal{T}_n$	Random training data $\{\boldsymbol{X}_i, i = 1, \ldots, n\}$.
g	Approximation of the pdf f.
$\text{Loss}(f(\boldsymbol{x}), g(\boldsymbol{x}))$	Loss incurred when approximating $f(\boldsymbol{x})$ with $g(\boldsymbol{x})$.
$\ell(g)$	Risk for approximation function g; that is, $\mathbb{E}\,\text{Loss}(f(\boldsymbol{X}), g(\boldsymbol{X}))$.
$g^{\mathcal{G}}$	Optimal approximation function in function class $\mathcal{G}$; that is, $\text{argmin}_{g \in \mathcal{G}}\, \ell(g)$.
$\ell_\tau(g)$	Training loss for approximation function (guess) g; that is, the sample average estimate of $\ell(g)$ based on a fixed training sample τ.
$\ell_{\mathcal{T}}(g)$	The same as $\ell_\tau(g)$, but now for a random training sample $\mathcal{T}$.
$g_\tau^{\mathcal{G}}$ or g_τ	The *learner*: $\text{argmin}_{g \in \mathcal{G}}\, \ell_\tau(g)$. That is, the optimal approximation function based on a fixed training set τ and function class $\mathcal{G}$. We suppress the superscript $\mathcal{G}$ if the function class is implicit.
$g_{\mathcal{T}}^{\mathcal{G}}$ or $g_{\mathcal{T}}$	The learner for a random training set $\mathcal{T}$.

Likelihood

A key step in setting up the learning procedure is to select a suitable function class $\mathcal{G}$ over which to optimize.

The standard approach is to parameterize g with a parameter $\boldsymbol{\theta}$ and let $\mathcal{G}$ be the class of functions $\{g(\cdot \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ for some p-dimensional parameter set Θ.

The function $\boldsymbol{\theta} \mapsto g(\boldsymbol{x} \mid \boldsymbol{\theta})$ is called the likelihood function. It gives the likelihood of the observed feature vector $\boldsymbol{x}$ under $g(\cdot \mid \boldsymbol{\theta})$, as a function of the parameter $\boldsymbol{\theta}$.

The natural logarithm of the likelihood function is called the log-likelihood function and its gradient with respect to $\boldsymbol{\theta}$ is called the score function, denoted $\boldsymbol{S}(\boldsymbol{x} \mid \boldsymbol{\theta})$; that is,

$$\boldsymbol{S}(\boldsymbol{x} \mid \boldsymbol{\theta}) := \frac{\partial \ln g(\boldsymbol{x} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\frac{\partial g(\boldsymbol{x} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}{g(\boldsymbol{x} \mid \boldsymbol{\theta})}.$$

Score

The random score $\boldsymbol{S}(\boldsymbol{X} \,|\, \boldsymbol{\theta})$, with $\boldsymbol{X} \sim g(\cdot \,|\, \boldsymbol{\theta})$, is of particular interest. In many cases, its expectation is *equal to the zero vector*; namely,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}} \boldsymbol{S}(\boldsymbol{X} \,|\, \boldsymbol{\theta}) &= \int \frac{\frac{\partial g(\boldsymbol{x} \,|\, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}{g(\boldsymbol{x} \,|\, \boldsymbol{\theta})} \, g(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{x} \\ &= \int \frac{\partial g(\boldsymbol{x} \,|\, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \, \mathrm{d}\boldsymbol{x} = \frac{\partial \int g(\boldsymbol{x} \,|\, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{x}}{\partial \boldsymbol{\theta}} = \frac{\partial 1}{\partial \boldsymbol{\theta}} = \boldsymbol{0},\end{aligned}$$

provided that the interchange of differentiation and integration is justified.

It is important to see whether expectations are taken with respect to $\boldsymbol{X} \sim g(\cdot \,|\, \boldsymbol{\theta})$ or $\boldsymbol{X} \sim f$. We use the expectation symbols $\mathbb{E}_{\boldsymbol{\theta}}$ and $\mathbb{E}$ to distinguish the two cases.

Fisher Information Matrix

The covariance matrix of the random score $\boldsymbol{S}(\boldsymbol{X}\,|\,\boldsymbol{\theta})$ is called the Fisher information matrix, which we denote by $\mathbf{F}$ or $\mathbf{F}(\boldsymbol{\theta})$ to show its dependence on $\boldsymbol{\theta}$. Since the expected score is $\mathbf{0}$, we have

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{S}(\boldsymbol{X}\,|\,\boldsymbol{\theta})\,\boldsymbol{S}(\boldsymbol{X}\,|\,\boldsymbol{\theta})^\top].$$

A related matrix is the expected Hessian matrix of $-\ln g(\boldsymbol{X}\,|\,\boldsymbol{\theta})$:

$$\mathbf{H}(\boldsymbol{\theta}) := \mathbb{E}\left[-\frac{\partial \boldsymbol{S}(\boldsymbol{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = -\mathbb{E}\begin{bmatrix} \frac{\partial^2 \ln g(\boldsymbol{X}\,|\boldsymbol{\theta})}{\partial^2 \theta_1} & \frac{\partial^2 \ln g(\boldsymbol{X}\,|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ln g(\boldsymbol{X}\,|\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2 \ln g(\boldsymbol{X}\,|\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ln g(\boldsymbol{X}\,|\boldsymbol{\theta})}{\partial^2 \theta_2} & \cdots & \frac{\partial^2 \ln g(\boldsymbol{X}\,|\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ln g(\boldsymbol{X}\,|\boldsymbol{\theta})}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 \ln g(\boldsymbol{X}\,|\boldsymbol{\theta})}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial^2 \ln g(\boldsymbol{X}\,|\boldsymbol{\theta})}{\partial^2 \theta_p} \end{bmatrix}$$

Note that the expectation here is with respect to $\boldsymbol{X} \sim f$.

Information Matrix Equality

It turns out that if $f = g(\cdot \mid \boldsymbol{\theta})$, the two matrices are the *same*; that is,

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbf{H}(\boldsymbol{\theta}), \tag{1}$$

provided that we may swap the order of differentiation and integration (expectation). This result is called the information matrix equality.

The matrices $\mathbf{F}(\boldsymbol{\theta})$ and $\mathbf{H}(\boldsymbol{\theta})$ play important roles in approximating the cross-entropy risk for large n.

To set the scene, let $g^{\mathcal{G}} = g(\cdot \mid \boldsymbol{\theta}^*)$ be the minimizer of the cross-entropy risk

$$r(\boldsymbol{\theta}) := -\mathbb{E} \ln g(\boldsymbol{X} \mid \boldsymbol{\theta}).$$

We assume that r, as a function of $\boldsymbol{\theta}$, is well-behaved; in particular, that in the neighborhood of $\boldsymbol{\theta}^*$ it is strictly convex and twice continuously differentiable (this holds true, for example, if g is a Gaussian density).

It follows that $\boldsymbol{\theta}^*$ is a root of $\mathbb{E}\,\boldsymbol{S}(\boldsymbol{X}\,|\,\boldsymbol{\theta})$, because

$$\mathbf{0} = \frac{\partial r(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} = -\frac{\partial\,\mathbb{E}\ln g(\boldsymbol{X}\,|\,\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} = -\mathbb{E}\frac{\partial \ln g(\boldsymbol{X}\,|\,\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} = -\mathbb{E}\,\boldsymbol{S}(\boldsymbol{X}\,|\,\boldsymbol{\theta}^*).$$

In the same way, $\mathbf{H}(\boldsymbol{\theta})$ is then the Hessian matrix of r.

Let $g(\cdot\,|\,\widehat{\boldsymbol{\theta}}_n)$ be the minimizer of the training loss

$$r_{\mathcal{T}_n}(\boldsymbol{\theta}) := -\frac{1}{n}\sum_{i=1}^{n}\ln g(\boldsymbol{X}_i\,|\,\boldsymbol{\theta}),$$

where $\mathcal{T}_n = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ is a training set. Let $r^* = -\mathbb{E}\ln f(\boldsymbol{X})$, where $\boldsymbol{X} \sim f$, be the smallest possible cross-entropy risk. We can decompose the generalization risk, $\ell(g(\cdot\,|\,\widehat{\boldsymbol{\theta}}_n)) = r(\widehat{\boldsymbol{\theta}}_n)$, into

$$r(\widehat{\boldsymbol{\theta}}_n) = r^* + \underbrace{r(\boldsymbol{\theta}^*) - r^*}_{\text{approx. error}} + \underbrace{r(\widehat{\boldsymbol{\theta}}_n) - r(\boldsymbol{\theta}^*)}_{\text{statistical error}}.$$

Cross-Entropy Risk for Large n

The following theorem specifies the asymptotic behavior of the components of the generalization risk. In the proof we assume that $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}^*$ as $n \to \infty$.

Theorem: Approximating the Cross-Entropy Risk

It holds asymptotically ($n \to \infty$) that

$$\mathbb{E}r(\widehat{\boldsymbol{\theta}}_n) - r(\boldsymbol{\theta}^*) \simeq \operatorname{tr}\left(\mathbf{F}(\boldsymbol{\theta}^*)\,\mathbf{H}^{-1}(\boldsymbol{\theta}^*)\right)/(2n), \tag{2}$$

where

$$r(\boldsymbol{\theta}^*) \simeq \mathbb{E}r_{\mathcal{T}_n}(\widehat{\boldsymbol{\theta}}_n) + \operatorname{tr}\left(\mathbf{F}(\boldsymbol{\theta}^*)\,\mathbf{H}^{-1}(\boldsymbol{\theta}^*)\right)/(2n). \tag{3}$$

Proof

A Taylor expansion of $r(\widehat{\boldsymbol{\theta}}_n)$ around $\boldsymbol{\theta}^*$ gives the statistical error

$$r(\widehat{\boldsymbol{\theta}}_n) - r(\boldsymbol{\theta}^*) = (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^\top \underbrace{\frac{\partial r(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}}}_{=\mathbf{0}} + \frac{1}{2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)^\top \mathbf{H}(\overline{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*), \quad (4)$$

where $\overline{\boldsymbol{\theta}}_n$ lies on the line segment between $\boldsymbol{\theta}^*$ and $\widehat{\boldsymbol{\theta}}_n$.

For large n we may replace $\mathbf{H}(\overline{\boldsymbol{\theta}}_n)$ with $\mathbf{H}(\boldsymbol{\theta}^*)$ as, by assumption, $\widehat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}^*$.

If $r_{\mathcal{T}}(\boldsymbol{\theta})$ is a differentiable function with respect to $\boldsymbol{\theta}$, then we can find the optimal parameter $\widehat{\boldsymbol{\theta}}_n$ by solving

$$\frac{\partial r_{\mathcal{T}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{S}(\boldsymbol{X}_i \mid \boldsymbol{\theta})}_{\boldsymbol{S}_{\mathcal{T}}(\boldsymbol{\theta})} = \mathbf{0}.$$

Proof (cont.)

In other words, the estimate $\widehat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$ is obtained by solving the root of the average score function, that is, by solving

$$\boldsymbol{S}_{\mathcal{T}}(\boldsymbol{\theta}) = \mathbf{0},$$

whereas $\boldsymbol{\theta}^*$ solves

$$\mathbb{E}\boldsymbol{S}(\boldsymbol{X} \,|\, \boldsymbol{\theta}) = \mathbf{0}.$$

We see thus that $\widehat{\boldsymbol{\theta}}_n$ is an M-estimator of $\boldsymbol{\theta}^*$.

Such estimators are asymptotically normal. In particular,

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\mathbf{0}, \mathbf{H}^{-1}(\boldsymbol{\theta}^*)\,\mathbf{F}(\boldsymbol{\theta}^*)\,\mathbf{H}^{-\top}(\boldsymbol{\theta}^*)\right). \qquad (5)$$

Proof (cont.)

Combining (4) with (5), the result (2) now follows from the following theorem (exercise!):

Theorem: Expectation of a Quadratic Form

Let $\mathbf{A}$ be an $n \times n$ matrix and $\boldsymbol{X}$ an n-dimensional random vector with expectation vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The random variable $Y := \boldsymbol{X}^\top \mathbf{A} \boldsymbol{X}$ has expectation $\operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu}$.

Next, we consider a Taylor expansion of $r_{\mathcal{T}_n}(\boldsymbol{\theta}^*)$ around $\widehat{\boldsymbol{\theta}}_n$:

$$r_{\mathcal{T}_n}(\boldsymbol{\theta}^*) = r_{\mathcal{T}_n}(\widehat{\boldsymbol{\theta}}_n) + (\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_n)^\top \underbrace{\frac{\partial r_{\mathcal{T}_n}(\widehat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}}_{=\mathbf{0}} + \frac{1}{2}(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_n)^\top \mathbf{H}_{\mathcal{T}_n}(\overline{\boldsymbol{\theta}}_n)(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_n), \quad (6)$$

where $\mathbf{H}_{\mathcal{T}_n}(\overline{\boldsymbol{\theta}}_n) := -\frac{1}{n}\sum_{i=1}^n \frac{\partial \boldsymbol{S}(\boldsymbol{X}_i \,|\, \overline{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}$ is the Hessian of $r_{\mathcal{T}_n}(\boldsymbol{\theta})$ at some $\overline{\boldsymbol{\theta}}_n$ between $\widehat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*$.

Proof (cont.)

Taking expectations on both sides of (6), we obtain

$$r(\boldsymbol{\theta}^*) = \mathbb{E} r_{\mathcal{T}_n}(\widehat{\boldsymbol{\theta}}_n) + \frac{1}{2}\mathbb{E}\,(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_n)^\top \mathbf{H}_{\mathcal{T}_n}(\overline{\boldsymbol{\theta}}_n)(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_n).$$

Replacing $\mathbf{H}_{\mathcal{T}_n}(\overline{\boldsymbol{\theta}}_n)$ with $\mathbf{H}(\boldsymbol{\theta}^*)$ for large n and using (5), we have

$$n\,\mathbb{E}\,(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_n)^\top \mathbf{H}_{\mathcal{T}_n}(\overline{\boldsymbol{\theta}}_n)(\boldsymbol{\theta}^* - \widehat{\boldsymbol{\theta}}_n) \longrightarrow \operatorname{tr}\left(\mathbf{F}(\boldsymbol{\theta}^*)\,\mathbf{H}^{-1}(\boldsymbol{\theta}^*)\right), \quad n \to \infty.$$

Therefore, asymptotically as $n \to \infty$, we have (3).

Consequences

The approximate Cross-Entropy risk theorem has a number of interesting consequences:

First, similar to the supervised learning case, the training loss $\ell_{\mathcal{T}_n}(g_{\mathcal{T}_n}) = r_{\mathcal{T}_n}(\widehat{\boldsymbol{\theta}}_n)$ tends to underestimate the risk $\ell(g^{\mathcal{G}}) = r(\boldsymbol{\theta}^*)$, because the training set $\mathcal{T}_n$ is used to both train $g \in \mathcal{G}$ (that is, estimate $\boldsymbol{\theta}^*$) and to estimate the risk.

The relation

$$r(\boldsymbol{\theta}^*) \simeq \mathbb{E} r_{\mathcal{T}_n}(\widehat{\boldsymbol{\theta}}_n) + \operatorname{tr}\left(\mathbf{F}(\boldsymbol{\theta}^*)\,\mathbf{H}^{-1}(\boldsymbol{\theta}^*)\right)/(2n).$$

tells us that on average the training loss underestimates the true risk by $\operatorname{tr}(\mathbf{F}(\boldsymbol{\theta}^*)\,\mathbf{H}^{-1}(\boldsymbol{\theta}^*))/(2n)$.

Consequences

Secondly, adding equations (2) and (3), yields the following asymptotic approximation to the expected generalization risk:

$$\mathbb{E}\, r(\widehat{\boldsymbol{\theta}}_n) \simeq \mathbb{E}\, r_{\mathcal{T}_n}(\widehat{\boldsymbol{\theta}}_n) + \frac{1}{n}\mathrm{tr}\left(\mathbf{F}(\boldsymbol{\theta}^*)\, \mathbf{H}^{-1}(\boldsymbol{\theta}^*)\right) \tag{7}$$

The first term on the right-hand side of (7) can be estimated (without bias) via the training loss $r_{\mathcal{T}_n}(\widehat{\boldsymbol{\theta}}_n)$.

For the second term, if the true model $f \in \mathcal{G}$, then $\mathbf{F}(\boldsymbol{\theta}^*) = \mathbf{H}(\boldsymbol{\theta}^*)$.

Therefore, for a p-dimensional vector $\boldsymbol{\theta}$, we may approximate the second term as $\mathrm{tr}(\mathbf{F}(\boldsymbol{\theta}^*)\mathbf{H}^{-1}(\boldsymbol{\theta}^*))/n \approx \mathrm{tr}(\mathbf{I}_p)/n = p/n$, giving:

$$\mathbb{E}\, r(\widehat{\boldsymbol{\theta}}_n) \approx r_{\mathcal{T}_n}(\widehat{\boldsymbol{\theta}}_n) + \frac{p}{n}. \tag{8}$$

AIC

Multiplying both sides of (7) by $2n$ and substituting $\operatorname{tr}\left(\mathbf{F}(\boldsymbol{\theta}^*)\mathbf{H}^{-1}(\boldsymbol{\theta}^*)\right) \approx p$, we obtain the approximation:

$$2n\, r(\widehat{\boldsymbol{\theta}}_n) \approx -2\sum_{i=1}^{n} \ln g(\boldsymbol{X}_i \,|\, \widehat{\boldsymbol{\theta}}_n) + 2p. \tag{9}$$

The right-hand side of (9) is called the Akaike information criterion (AIC).

Just like (8), the AIC approximation can be used to compare the difference in generalization risk of two or more learners. We prefer the learner with the smallest (estimated) generalization risk.

Maximum Likelihood Estimate

We have seen that parameter $\widehat{\boldsymbol{\theta}}_n$ that minimizes the training loss is found by solving

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{S}(\boldsymbol{X}_i \,|\, \boldsymbol{\theta})}_{\boldsymbol{S}_{\mathcal{T}}(\boldsymbol{\theta})} = \mathbf{0}. \tag{10}$$

This $\widehat{\boldsymbol{\theta}}_n$ is the maximum likelihood estimate of $\boldsymbol{\theta}$, as it minimizes

$$-\frac{1}{n}\sum_{i=1}^{n} \ln g(\boldsymbol{X}_i \,|\, \boldsymbol{\theta}) \quad \text{and hence maximizes} \quad \prod_{i=1}^{n} g(X_i \,|\, \boldsymbol{\theta}).$$

It is often not possible to find $\widehat{\boldsymbol{\theta}}$ in an explicit form. In that case one needs to solve the equation (10) numerically.

A standard technique for root-finding is Newton's method.

Newton's Method

Here, starting from an initial guess $\boldsymbol{\theta}_0$, subsequent iterates are obtained via the iterative scheme

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \mathbf{H}_{\mathcal{T}}^{-1}(\boldsymbol{\theta}_t)\, \boldsymbol{S}_{\mathcal{T}}(\boldsymbol{\theta}_t),$$

where

$$\mathbf{H}_{\mathcal{T}}(\boldsymbol{\theta}) := \frac{-\,\partial \boldsymbol{S}_{\mathcal{T}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial \boldsymbol{S}(\boldsymbol{X}_i \,|\, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

is the average Hessian matrix of $\{-\ln g(\boldsymbol{X}_i \,|\, \boldsymbol{\theta})\}_{i=1}^{n}$.

Under $f = g(\cdot \,|\, \boldsymbol{\theta})$, the expectation of $\mathbf{H}_{\mathcal{T}}(\boldsymbol{\theta})$ is equal to the information matrix $\mathbf{F}(\boldsymbol{\theta})$, which does not depend on the data.

This suggests an alternative iterative scheme, called Fisher's scoring method:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \mathbf{F}^{-1}(\boldsymbol{\theta}_t)\, \boldsymbol{S}_{\mathcal{T}}(\boldsymbol{\theta}_t), \tag{11}$$

which is not only easier to implement (if the information matrix can be readily evaluated), but also is more numerically stable.

Maximum Likelihood for the Gamma Distribution

We wish to approximate the density of the $\mathsf{Gamma}(\alpha^*, \lambda^*)$ distribution for some true but unknown parameters α^* and λ^*, on the basis of a training set $\tau = \{x_1, \ldots, x_n\}$ of iid samples from this distribution.

Choosing our approximating function $g(\cdot \,|\, \alpha, \lambda)$ in the same class of gamma densities,

$$g(x \,|\, \alpha, \lambda) = \frac{\lambda^\alpha x^{\alpha-1} \mathrm{e}^{-\lambda x}}{\Gamma(\alpha)}, \quad x \geqslant 0, \tag{12}$$

with $\alpha > 0$ and $\lambda > 0$, we seek to solve (10).

Taking the logarithm in (12), the log-likelihood function is given by

$$l(x \,|\, \alpha, \lambda) := \alpha \ln \lambda - \ln \Gamma(\alpha) + (\alpha - 1) \ln x - \lambda x.$$

It follows that

$$\boldsymbol{S}(\alpha,\lambda) = \begin{bmatrix} \frac{\partial}{\partial\alpha} l(x \,|\, \alpha,\lambda) \\ \frac{\partial}{\partial\lambda} l(x \,|\, \alpha,\lambda) \end{bmatrix} = \begin{bmatrix} \ln\lambda - \psi(\alpha) + \ln x \\ \frac{\alpha}{\lambda} - x \end{bmatrix},$$

where ψ is the derivative of $\ln\Gamma$: the so-called digamma function.

Hence,

$$\mathbf{H}(\alpha,\lambda) = -\mathbb{E}\begin{bmatrix} \frac{\partial^2}{\partial\alpha^2} l(X \mid \alpha,\lambda) & \frac{\partial^2}{\partial\alpha\partial\lambda} l(X \mid \alpha,\lambda) \\ \frac{\partial^2}{\partial\alpha\partial\lambda} l(X \mid \alpha,\lambda) & \frac{\partial^2}{\partial\lambda^2} l(X \mid \alpha,\lambda) \end{bmatrix} = -\mathbb{E}\begin{bmatrix} -\psi'(\alpha) & \frac{1}{\lambda} \\ \frac{1}{\lambda} & -\frac{\alpha}{\lambda^2} \end{bmatrix} = \begin{bmatrix} \psi'(\alpha) & -\frac{1}{\lambda} \\ -\frac{1}{\lambda} & \frac{\alpha}{\lambda^2} \end{bmatrix}.$$

Fisher's scoring method (11) can now be used to solve (10), with

$$\boldsymbol{S}_\tau(\alpha,\lambda) = \begin{bmatrix} \ln\lambda - \psi(\alpha) + n^{-1}\sum_{i=1}^n \ln x_i \\ \frac{\alpha}{\lambda} - n^{-1}\sum_{i=1}^n x_i \end{bmatrix}$$

and $\mathbf{F}(\alpha,\lambda) = \mathbf{H}(\alpha,\lambda)$.