7

Prediction and Classification

7.1 Prediction and loss functions

If (X, Y) denote a pair of random variables, we are interested in *predicting* the value of Y given that we have observed $X = x$. In general, the sample space of X and Y may be arbitrary, E_1 and E_2, say, and their simultaneous distribution on $E_1 \times E_2$ will be denoted P. The marginal distribution of X and Y will be denoted P_1 and P_2, respectively. If X and Y are independent, there is no point in predicting Y on the basis of observing X. Indeed, conditionally on $X = x$ the distribution of Y would be P_2 irrespectively of the value of $x \in E_1$.

From a probabilistic point of view, all that we need to do is to compute the conditional distribution of Y given $X = x$. Then this probability measure tells us exactly which values Y can attain with high probability if $X = x$. The conditional probability measure is, however, often a little complicated to interpret, nor is it really what we want to report as a "prediction". As a prediction we want to report an element in the sample space E_2.

If the sample spaces are both subsets of $\mathbb{R}^n$ or alternatively discrete, and if P is given by the density (or point probabilities) $f : E_1 \times E_2 \to (0, \infty)$, then the conditional distribution of Y given $X = x$ can be specified by the conditional density (or conditional point probabilities)

$$f(y|x) = \frac{f(x, y)}{f_1(x)} \tag{7.1}$$

where $f_1(x) = \int_{E_2} f(x, y)\mathrm{d}y$ is the density for the marginal distribution of X. If the sample spaces are both discrete, the marginal point probabilities for X are given by $f_1(x) = \sum_{y \in E_2} f(x, y)$.

An important case that we are going to consider in more details in the next section is where E_2 is discrete and E_1 in principle can be an arbitrary set. Then it is

often the case that the simultaneous distribution of (X, Y) is given by the marginal distribution of Y (in terms of point probabilities f_2) and the conditional distribution of X given $Y = y$. If $E_1 \subseteq \mathbb{R}^n$ and the conditional distribution has density $g(x|y)$, we find that the marginal distribution of X is given by

$$\begin{aligned}
\mathbb{P}(X \in A) &= \sum_{y \in E_2} \mathbb{P}(X \in A, Y = y) \\
&= \sum_{y \in E_2} \mathbb{P}(X \in A | Y = y) f_2(y) \\
&= \sum_{y \in E_2} \int_A g(x|y) \mathrm{d}x f_2(y) \\
&= \int_A \sum_{y \in E_2} f_2(y) g(x|y) \mathrm{d}x
\end{aligned}$$

which shows that the marginal distribution of X has density

$$f_1(x) = \sum_{y \in E_2} f_2(y) g(x|y).$$

The conditional distribution of Y given $X = x$ is then given by the point probabilities

$$f(y|x) = \frac{f_2(y) g(x|y)}{f_1(x)}. \tag{7.2}$$

You may check that they are positive and sum to one by the definition of f_1.

In situations like (7.1) or (7.2) where we have given the conditional distribution of Y given $X = x$ either in terms of a density or in terms of point probabilities, a sensible choice of *predictor* is the *maximum a posteriori predictor*.

$$\hat{y}(x) = \arg\max_y f(y|x).$$

Especially in the case with discrete E_2 sample space we can make sense out of this predictor, as it produces the single element in E_2 that have the highest probability in the conditional distribution of Y given $X = x$.

We observe that to use the maximum posteriori predictor, we do not need to carry out the computation of $f_1(x)$, since for given x this is a constant that does not depend upon y. There is another structurally interesting observation. We rewrite the joint density in (7.1) as $f(x, y) = f_2(y) g(x|y)$ where f_2 denotes the marginal density for the distribution of Y and $g(x|y)$ the conditional density for the distribution of X given $Y = y$. Then the *maximum a posteriori predictor* in both cases considered above can be computed as

$$\hat{y}(x) = \arg\max_y f_2(y) g(x|y) = \arg\max_y \log f_2(y) + \log g(x|y).$$

The last term here, $\log g(x|y)$, is structurally similar to the logarithm of a likelihood function given the observation $x \in E_1$, and we maximize over the unknown (unobserved) y. The other term, $\log f_2(y)$, depends upon y only – and thus doesn't change for different observations x – and it can be understood as regulating the prediction in concordance with the marginal distribution of Y.

Example 7.1.1. If $E_2 = \mathbb{R}$ and the conditional distribution of Y given $X = x$ is a normal distribution with mean value $g(x)$ that depends upon x through the function $g : E_1 \to \mathbb{R}$, and variance $\sigma^2 > 0$, then

$$f(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - g(x))^2}{2\sigma^2}\right).$$

Thus the maximum a posteriori predictor is

$$\arg\min_{y\in\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - g(x))^2}{2\sigma^2}\right) = \arg\min_{y\in\mathbb{R}} -(y - g(x))^2 = g(x).$$

Thus the maximum a posteriori predictor is the conditional expectation of Y given $X = x$. ◇

It is natural to ask what properties a good predictor should have? But is not so easy to answer this question. The attempt to do so will always depend upon the context and the purpose for which the predictor is going to be used. Are we trying to predict the future development of stock prices given the history, are we making a diagnose for a patient based on a series of test result, or are we making computational annotations of biological sequences. The applications of predictions span a huge number of different subjects and problems. The approach to unify the theory for choosing good predictors leads to *statistical decision theory*. This is based on the introduction of a *loss function* $L : E_2 \times E_2 \to [0, \infty)$, such that if $\hat{y} : E_1 \to E_2$ is any predictor, then

$$L(y, \hat{y}(x))$$

is the loss of predicting $\hat{y}(x)$ when $Y = y$. A big loss is not desirable. The distribution of the random variable $L(Y, \hat{y}(X))$, which is given in terms of a transformation of the simultaneous distribution of (X, Y), is precisely the distribution of the losses that we will suffer when using the predictor $\hat{y}$. To compare predictors using the loss function L, we should therefore compare their corresponding loss-distributions. The closer to 0 the loss-distribution is located the better.

Since it is difficult to compare distributions, one often rely on comparing their mean values. The *expected prediction error* for the predictor $\hat{y}$ is defined as

$$\mathrm{EPE}(\hat{y}) = \mathbb{E}(L(Y, \hat{y}(X))).$$

The optimal predictor is then defined as the predictor that minimizes EPE over the set of possible predictors.

There are two important observations. Since the quality of the predictor is measured by the loss function L, the optimal predictor is only optimal in the sense that L dictates. Different choices of loss functions lead to different optimal solutions. It is in general never possible to come up with a predictor that is optimal in every thinkable way – only in the the sense of minimizing the expected loss for a particular loss function. The other observation is that we consider only the expectation of $L(Y, \hat{y}(X))$ as a measure good performance, when we use EPE.

Example 7.1.2. The standard loss function is the *squared error loss*. When $E_2 \subseteq \mathbb{R}$ we define the squared error loss function as

$$L_2(y, \hat{y}) = (y - \hat{y})^2.$$

With the distribution of X having density f_1 we find that

$$\begin{aligned} \mathrm{EPE}(\hat{y}) &= \mathbb{E}((\hat{y}(X) - Y)^2) \\ &= \int\int (\hat{y}(x) - y)^2 f_1(x) f(y|x) \mathrm{d}x \mathrm{d}y \\ &= \int\int (\hat{y}(x) - y)^2 f(y|x) \mathrm{d}y f_1(x) \mathrm{d}x \end{aligned}$$

We find that if we choose $\hat{y}(x)$ to minimize the inner integral

$$\int (\hat{y}(x) - y)^2 f(y|x) \mathrm{d}y$$

for every $x \in E_1$, we have also minimized EPE. We recognize this integral as the expectation of $(\hat{y}(x) - Y)^2$ in the conditional distribution of Y given $X = x$, which we will write as $\mathbb{E}((\hat{y}(x) - Y)^2 | X = x)$. By (4.11) we have that

$$\mathbb{E}((\hat{y}(x) - Y)^2 | X = x) = \mathbb{V}(Y|X = x) + (\hat{y}(x) - \mathbb{E}(Y|X = x))^2,$$

which is minimized if and only if $\hat{y}(x) = \mathbb{E}(Y|X = x)$. That is, the optimal predictor in the sense of minimizing the expected prediction error when using the squared error loss is given as

$$\hat{y}(x) = \mathbb{E}(Y|X = x) = \int y f(y|x) \mathrm{d}y$$

where $\mathbb{E}(Y|X = x)$ is the conditional expectation of Y given that $X = x$.

If $E_2 \subseteq \mathbb{R}^n$ a simple generalization of the squared error loss to n-dimensional vectors is obtained by taking

$$L(y, \hat{y}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

which is the sum of squared deviations for each of the coordinates. It is not difficult to see that the argument above can be carried out coordinate by coordinate to give the optimal predictor

$$\hat{y}_i(x) = \mathbb{E}(Y_i|X = x).$$

$\diamond$

Example 7.1.3. In Example 2.8.17 we considered the linear regression model, where the distribution of (X, Y) on $\mathbb{R}^2$ was given in terms of the marginal distribution of X and the conditional distribution of Y given X. In particular we have that the conditional mean of Y given $X = x$ was specified to be

$$\mathbb{E}(Y|X = x) = \alpha + \beta x.$$

Thus the optimal predictor in the sense of squared error loss of Y given $X = x$ is simply $\hat{y}(x) = \alpha + \beta x$. ◇

Example 7.1.4. Another possible choice of loss function is the absolute value of the deviation instead of the squared deviation. That is,

$$L_1(y, \hat{y}) = |y - \hat{y}|.$$

The solution is that the optimal predictor is the *conditional median* of Y in the conditional distribution of Y given $X = x$.

Depending upon a parameter $p > 0$ we can choose a range of different loss functions,

$$L_p(y, \hat{y}) = |y - \hat{y}|^p.$$

The tradeoff is that small values of p put more emphasis on the numerically small prediction errors whereas large values of p put more emphasis of the large prediction errors. ◇

The discussion about how to choose the proper predictor is by no means settled by the abstract decision theory, where we introduce the loss function to provide a way of quantifying different properties of different predictors. We have simply moved the problem from theory to practice. We will in practice need to specify the loss function that captures precisely the needs of that particular problem before we rely on the theory. This is not so easy – how do we quantify a "loss" in situations where there is no obvious or direct financial loss by a wrong prediction. Second, even if we manage to come up with the "right" loss function, it may not really be easy to work with in the computation of the optimal predictor. The squared error loss, L_2, is so convenient to work with that it is often used.

7.2 Classification

When we predict a variable Y that take discrete values, we often talk about classification instead. This is in particular the case if Y can take only finitely many different values, that can be regarded as some kind of group label attached to the X-variable. Thus we then classify the observation x to one of the groups by predicting the group label. The special case with E_2 containing only two different elements, and where we often take $E_2 = \{0, 1\}$, plays a central role in the theory as well as in many applications.

For the finite classification problem we can in principle rather easily write down all possible loss functions, as we will do in the following example.

Example 7.2.1. If $E_2 = \{1, \ldots, n\}$ is a discrete sample space containing n elements, the values, $L(y, \hat{y})$, of any loss function can be viewed and organized as a matrix of weights;

$$L = \begin{pmatrix} L(1,1) & L(1,2) & \ldots & L(1,n) \\ \vdots & & & \vdots \\ L(n,1) & L(n,2) & \ldots & L(n,n) \end{pmatrix}.$$

Usually we take the diagonal elements in the L matrix to be 0, that is, there is no loss when the prediction is correct, and the off-diagonals to be > 0.

A special case is obtained by $L(i,j) = 1$ whenever $i \neq j$ and $L(i,i) = 0$. This is often called the *zero-one* loss function, as it gives a loss of 1 whenever we make a wrong prediction and a loss of 0 whenever we make a correct prediction. We find that

$$\begin{aligned} \text{EPE}(\hat{y}) &= \mathbb{E}(L(Y, \hat{y}(X))) \\ &= \int \left(\sum_y L(y, \hat{y}(x))) f(y|x) \right) f_1(x) \mathrm{d}x. \end{aligned}$$

If we use the zero-one loss function, $L(Y, \hat{y}(X))$ is a Bernoulli variable, and it's expectation, $\text{EPE}(\hat{y})$, equals the probability that the loss is 1, that is, the expected prediction error for the zero-one loss function is the probability of misclassification.

The expression for $\text{EPE}(\hat{y})$ is minimized by minimizing the inner sum for every $x \in E_1$. When we use the zero-one loss function this sum is

$$\begin{aligned} \sum_y L(y, \hat{y}(x))) f(y|x) &= \sum_{y \neq \hat{y}(x)} f(y|x) \\ &= 1 - f(\hat{y}(x)|x). \end{aligned}$$

The optimal choice of $\hat{y}$, that minimizes $1 - f(\hat{y}(x)|x)$ for every $x \in E_1$, is therefore

$$\hat{y}(x) = \arg\max_y f(y|x),$$

which is precisely the maximum a posteriori predictor. This is also known as the *Bayes classifier*, and the expected prediction error for the Bayes classifier is called the Bayes rate. We see that the Bayes rate is the probability of misclassification when using the Bayes classifier. ◇

With a dichotomous Y-variable, $E_2 = \{0, 1\}$, we typically change the notation for the marginal distribution of Y, and write π_0 and π_1 (with $\pi_0 + \pi_1 = 1$) for the point probabilities $f_2(0)$ and $f_2(1)$. Then the Bayes classifier is given by computing the maximum of $\pi_0 g(x|0)$ and $\pi_1 g(x|1)$. We see that this divides the sample space E_1 into two sets

$$G_0 = \left\{ x \in E_1 \;\middle|\; \frac{g(x|0)}{g(x|1)} > \frac{\pi_1}{\pi_0} \right\}$$

and

$$G_1 = \left\{ x \in E_1 \;\middle|\; \frac{g(x|0)}{g(x|1)} < \frac{\pi_1}{\pi_0} \right\},$$

such that the Bayes classifier $\hat{y}$ equals 0 on G_0 and 1 on G_1. On the boundary between the two sets where we have equality, we can predict 0 or 1 as we like. We see the effect of changing the proportions π_0 and π_1. If $\pi_0 = \pi_1 = 0.5$ we classify according to y-variable which gives the the largest conditional probability of the observed x. As π_1 gets larger the set G_0 shrinks and the set G_1 grows accordingly, and vice versa, if π_1 gets smaller G_0 grows and G_1 shrinks accordingly.

To summarize the performance of any classification procedure we can benefit from the *confusion matrix*:

	Predicted y	
Observed y	0	1
0	p_{00}	p_{01}
1	p_{10}	p_{11}

Table 7.1: The confusion matrix gives the probability of making different classifications divided according to the observed value of Y. The sum $p_{01} + p_{10}$ is the probability of misclassification, which is always bounded from below by the Bayes rate.

The off-diagonal elements in the confusion matrix represents the two different types of misclassifications. Either the observed y equals 0 and we classify as 1 or it equals 1 and we classify as 0. The zero-one loss function puts equal weight on the two types of mistakes and the Bayes classifier minimizes the sum $p_{01} + p_{10}$. Changing the off diagonal elements $L(0,1)$ and $L(1,0)$ in the loss function, so that they differ, we put different weights on the two types of misclassifications, and the optimal predictor will minimize

$$L(0,1)p_{01} + L(1,0)p_{10}$$

instead. It may, for instance, be much more serious to make the mistake of diagnosing a patient to be healthy, even though the patient is ill than the other way around.

As always there is here a tradeoff, and a reasonable tradeoff can only be decided upon in the concrete situation.

7.3 Estimation and expected loss

In the previous sections we derived several optimal procedures for prediction or classification. The optimality is in the sense that the procedure gives the smallest expected prediction error – or in other words the smallest expected loss – for a particular choice of loss function. However, the constructions rely on complete knowledge

of the simultaneous distribution of (X, Y). This is something we never have, but something that we can try to infer (estimate) from a dataset.

The ideal situation is that we have a convenient, parameterized family P_θ, $\theta \in \Theta$, of probability measures on $E_1 \times E_2$ such that the distribution of (X, Y) belongs to this family, and that we have a dataset consisting of n iid realizations $(x_1, y_1), \ldots, (x_n, y_n)$ of (X, Y). Then we can simply estimate θ and compute the optimal predictor for the estimated probability measure.

Example 7.3.1. Again considering Example 2.1.17, the probability measure on $\mathbb{R}^2$ is parameterized by $\alpha, \beta \in \mathbb{R}$ and the two variance parameters $\sigma_1^2, \sigma_2^2 > 0$. Using least squared regression we can estimate α and β from a dataset to obtain the estimates $\hat{\alpha}$ and $\hat{\beta}$. Then, according to Example 7.1.3, the estimated optimal predictor is given as

$$\hat{y}(x) = \hat{\alpha} + \hat{\beta}x.$$

$\diamond$

Example 7.3.2 (Mixtures)**.** If $E_1 = \mathbb{R}$ and the conditional distribution of X given $Y = y$ is given by the distribution function G_y, then

$$P(X \leq x, Y = y) = \pi_y G_y(x)$$

and the marginal distribution of X has distribution function

$$P(X \leq x) = \pi_0 G_0(x) + \pi_1 G_1(x).$$

If G_y is differentiable, and thus the conditional distribution is given by a density, $g(x|y)$, say, then the marginal distribution of X has density

$$f_1(x) = \pi_0 g(x|0) + \pi_1 g(x|1).$$

In general, we will estimate π_0 (and implicitly π_1) as the relative frequencies of $y_i = 0$ in the dataset. The conditional densities can be estimated based on the relevant x_i-observations where either $y_i = 0$ or $y_i = 1$.

If both conditional distributions are normal with the same variance, that is,

$$g(x|y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu(y))^2}{2\sigma^2}\right),$$

then

$$\begin{aligned} \frac{g(x|0)}{g(x|1)} &= \exp\left(-\frac{\mu(0)^2 - 2\mu(0)x + 2\mu(1)x - \mu(1)^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{2(\mu(1) - \mu(0))x + \mu(0)^2 - \mu(1)^2}{2\sigma^2}\right). \end{aligned}$$

By taking logarithms we see that $\pi_0 g(x|0) = \pi_1 g(x|1)$ if and only if

$$x = \frac{2\sigma^2(\log \pi_1 - \log \pi_0) + \mu(1)^2 - \mu(0)^2}{2(\mu(1) - \mu(0))}.$$

For the special case $\pi_0 = \pi_1 = 0.5$ this simplifies to

$$x = \frac{\mu(0) + \mu(1)}{2}$$

where we used that $\mu(1)^2 - \mu(0)^2 = (\mu(1) - \mu(0))(\mu(1) + \mu(0))$. This boundary point between the two sets G_1 and G_0 is precisely the midpoint between the two means. If $\pi_0 \neq 0.5$ the boundary point moves towards the mean for which π_y is smallest.

In general, there may not be a unique boundary point and the sets G_0 and G_1 may not be simple intervals. This happens for instance if the two normal densities above are allowed to have different variance. But often there will in reality only be one boundary point that is interesting. Others will typically be out in the tail of the densities where it is highly unlikely that we will ever get observations. ◇

Example 7.3.3. In Example 5.2.18 we considered the logistic regression model. In this case we would be interested in predicting the dichotomous variable death/survival of a fly for a given concentration of dimethoat. This is an example where we naturally have a model of the conditional distribution of Y given $X = x$ directly – rather than a model of the simultaneous distribution of (X, Y). The parameters are $\alpha, \beta \in \mathbb{R}$, and based on a dataset, we can obtain the estimates $\hat{\alpha}$ and $\hat{\beta}$. The estimated point probability $\hat{p}(x)$ of $Y = 1$ given x is

$$\hat{p}(x) = \frac{\exp(\hat{\alpha} + \hat{\beta}x)}{1 + \exp(\hat{\alpha} + \hat{\beta}x)}.$$

Since the logit transformation $p \to \log(p/(1-p))$ is monotonely increasing, we find that $\hat{p}(x) > 1/2$ if and only if $\hat{\alpha} + \hat{\beta}x > 0$, and the Bayes classifier is therefore given as

$$\hat{y}(x) = \begin{cases} 1 & \text{if } \hat{\alpha} + \hat{\beta}x > 0 \\ 0 & \text{if } \hat{\alpha} + \hat{\beta}x < 0 \end{cases}.$$

In particular we observe that LD_{50} is the boundary point that separates the G_0 set from the G_1 set when $\pi_0 = \pi_1$. ◇

Whenever we estimate the parameters, the optimality of the resulting estimated predictor is lost. The estimated predictor is optimal for the *estimated* model, but not the true model. We are interested in finding out how much we have lost.

If $\hat{y}_\theta$ denotes the optimal predictor for P_θ, then

$$\text{EPE}(\theta) = \mathbb{E}_\theta(Y, L(\hat{y}_\theta(X)))$$

is the minimal expected prediction error we can obtain for $\theta \in \Theta$. We can regard $\text{EPE}(\theta)$ as a *parameter of interest*, we can estimate the optimal EPE as $\text{EPE}(\hat{\theta})$, and we may use bootstrapping for instance to give a 95% confidence interval for $\text{EPE}(\theta)$. But this is not really what we want! It is fine that we can tell that the optimal predictor has an EPE that is in this and that interval, but we don't know what the optimal predictor looks like exactly. What we have is instead an estimated predictor $\hat{y}_{\hat{\theta}}$, and we would like to know the size of the expected loss for that predictor. Therefore we introduce another quantity

$$\text{Err}(\theta) = \mathbb{E}_\theta(L(Y, \hat{y}_{\hat{\theta}}(X))),$$

where the expectation $\mathbb{E}_\theta$ denotes expectation over the estimator $\hat{\theta}$ as well as an *independent* copy of (X, Y). The quantity $\text{Err}(\theta)$ as a function of θ is called the *test error* or *generalization error*. It is important to understand the definition correctly. If θ is the true and fixed parameter, the estimator $\hat{\theta}$ has a distribution. When we apply the estimated parameter value to produce the estimated predictor $\hat{y}_{\hat{\theta}}$ and then predict another independent Y given X – still with θ the true parameter – the loss is $L(Y, \hat{y}_{\hat{\theta}}(X))$, and $\text{Err}(\theta)$ is the expectation of these losses over the distribution of the estimator as well as the independent (X, Y).

We can again regard $\text{Err}(\theta)$ as a parameter of interest, and we can estimate it as $\text{Err}(\hat{\theta})$ and compute 95% confidence intervals. The actual computation of the expectation in the definition of $\text{Err}(\theta)$ can be carried out by simulation, if we can not do it analytically (which we most likely can not).

We are still not completely satisfied. If we introduce the notation $\mathbb{E}_\theta^{\hat{\theta}}$ to denote the expectation operator where we take expectation w.r.t. the distribution of $\hat{\theta}$ and $\mathbb{E}_\theta^{(X,Y)}$ the expectation operator where we take expectation w.r.t. the distribution of (X, Y), we can rewrite

$$\text{Err}(\theta) = \mathbb{E}_\theta^{\hat{\theta}} \mathbb{E}_\theta^{(X,Y)}(L(Y, \hat{y}_{\hat{\theta}}(X))).$$

We are in reality interested in the inner expectation

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta^{(X,Y)}(L(\hat{y}_{\hat{\theta}}(X), Y)) \geq \mathbb{E}_\theta^{(X,Y)}(L(\hat{y}_\theta(X), Y))).$$

The inequality follows from the fact that $\hat{y}_\theta$ is the optimal predictor under P_θ. If $\hat{\vartheta}$ denotes our estimate of θ based on a dataset, the quantity $R(\hat{\vartheta}, \theta)$ is the expected prediction error that is of interest. This is a realization of the random variable $R(\hat{\theta}, \theta)$, but since we do not know θ we can not compute the quantity, and if we plug in the estimate of θ we get

$$R(\hat{\vartheta}, \hat{\vartheta}) = \text{EPE}(\hat{\vartheta}).$$

If $\hat{\vartheta}$ is close to θ we have almost equality in the inequality above but the size of the gap depends upon the concrete realization of $\hat{\vartheta}$. The quantity $\text{Err}(\theta)$ represents the expectation of the interesting quantity $R(\hat{\theta}, \theta)$ – that is, the average expected loss for the estimated predictor.

To really bound the expected loss, $R(\hat{\vartheta}, \theta)$, for the estimated predictor, we can introduce yet another quantity

$$z_q(\theta) = q\text{-quantile for the distribution of } R(\hat{\theta}, \theta) \text{ under } P_\theta.$$

Taking $q = 0.95$ we get that with 95% probability the expected loss of the estimated predictor is $\leq z_{0.95}(\theta)$. Again, regarding this as the parameter of interest, we can estimate this quantile as $z_{0.95}(\hat{\vartheta})$ and compute a 95% confidence interval for this quantile, to give the ultimate upper bound on the expected loss of the estimated predictor. Just as for the computation of $\mathrm{Err}(\theta)$ we will most likely have to use simulations to actually compute $z_{0.95}(\hat{\vartheta})$. Note that to simulate a single realization of $R(\hat{\theta}, \hat{\vartheta})$ we have to simulate an entire dataset using $P_{\hat{\vartheta}}$ (of exactly the same size as the original dataset), reestimate θ as $\hat{\vartheta}_{\text{new}}$, simulate a number of independent variables, again using $P_{\hat{\vartheta}}$, and make the prediction using the *reestimated* predictor based on $\hat{\vartheta}_{\text{new}}$, and finally compute the loss for each prediction. Then we take the average of these losses as an approximation to $R(\hat{\vartheta}_{\text{new}}, \hat{\vartheta})$. This gives a single realization and must be repeated a sufficiently large number of times to compute $z_{0.95}(\hat{\vartheta})$, say. If you use bootstrapping around all this there is yet another outer level of replications.

Algorithm 7.3.4 (Plug-in estimate of Err)**.** We consider $\mathrm{Err}(\theta) = \mathbb{E}_\theta^{\hat{\theta}} R(\hat{\theta}, \theta)$ and want to compute the plug-in estimate $\mathrm{Err}(\hat{\vartheta})$.

- Choose B sufficiently large and simulate B new independent, identically distributed datasets, $(x_1, y_1), \ldots, (x_B, y_B) \in E$, each simulation being from the probability measure $P_{\hat{\theta}}$.
- Compute, for each dataset (x_i, y_i), $i = 1, \ldots, B$, new estimates $\hat{\vartheta}_i = \hat{\theta}(x_i, y_i)$ using the estimator $\hat{\theta}$ and new optimal predictors $\hat{y}_{\hat{\vartheta}_i}$.
- Compute $R(\hat{\vartheta}_i, \hat{\vartheta})$ for $i = 1, \ldots, B$. This may again be done via simulations.
- Compute

$$\hat{\mathrm{Err}}(\hat{\vartheta}) = \frac{1}{B} \sum_{i=1}^{B} R(\hat{\vartheta}_i, \hat{\vartheta}).$$

By taking B suitably large, $\hat{\mathrm{Err}}(\hat{\vartheta})$ is an arbitrarily good approximation of $\mathrm{Err}(\hat{\vartheta})$. A similar algorithm can be used to compute $z_q(\hat{\vartheta})$, say, by changing the last step to a computation of the empirical q-quantile instead of an average.

A possible variation on the first two steps is to fix the y-variables as observed and simulate the X-variables conditionally on the y's.

How to estimate θ? Often one relies on methods such as MLE, but if the sole purpose of estimating θ is prediction, and if we have a loss function that we are really keen on optimizing, then it is a sound principle to try to estimate the parameters by optimizing the empirical loss. That is, we let $l(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_\theta(x_i))$ for $\theta \in$

Θ. We do not need a generative probability model behind this approach, just a parameterized family of predictors $\hat{y}_\theta$. Then we would estimate θ by

$$\hat{\theta} = \arg\min l(\theta).$$

The squared error loss leads for instance to least squares estimation. The zero-one loss is a little difficult to use directly as there will often be many optimal θ-values and the optimization problem will be very difficult as the function, $l(\theta)$, will be discontinuous and quite ugly.