# 序列模型

想象一下你正在看网飞（Netflix，一个国外的视频网站）上的电影。 作为一名忠实的用户，你对每一部电影都给出评价， 毕竟一部好电影需要更多的支持和认可。 然而事实证明，事情并不那么简单。 随着时间的推移，人们对电影的看法会发生很大的变化。 事实上，心理学家甚至对这些现象起了名字：

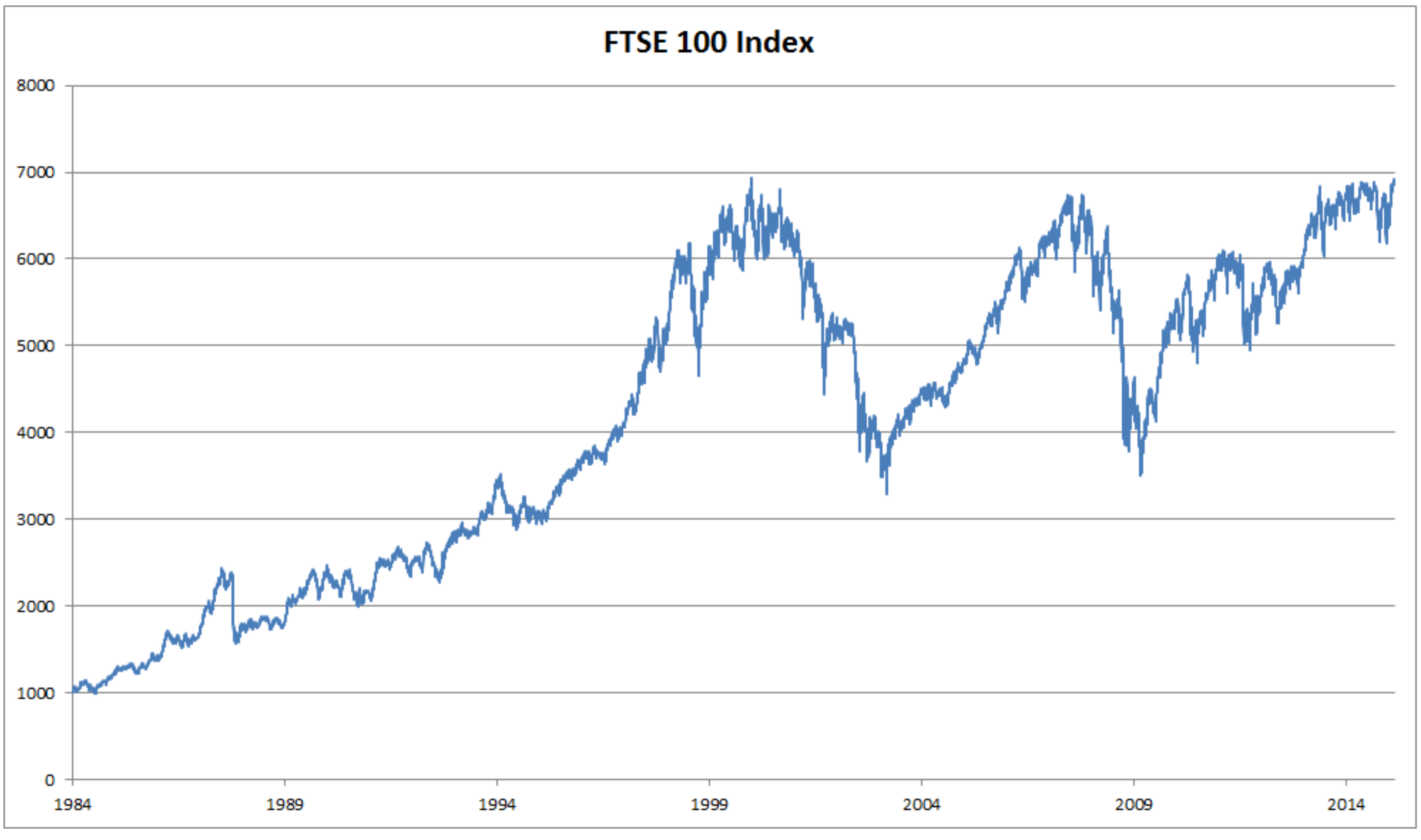
- *锚定*（anchoring）效应：基于其他人的意见做出评价。 例如，奥斯卡颁奖后，受到关注的电影的评分会上升，尽管它还是原来那部电影。 这种影响将持续几个月，直到人们忘记了这部电影曾经获得的奖项。 结果表明，这种效应会使评分提高半个百分点以上。
- *享乐适应*（hedonic adaption）： 人们迅速接受并且适应一种更好或者更坏的情况 作为新的常态。 例如，在看了很多好电影之后，人们会强烈期望下部电影会更好。 因此，在许多精彩的电影被看过之后，即使是一部普通的也可能被认为是糟糕的。
- *季节性*（seasonality）： 少有观众喜欢在八月看圣诞老人的电影。
- 有时，电影会由于导演或演员在制作中的不当行为变得不受欢迎。
- 有些电影因为其极度糟糕只能成为小众电影。*Plan9from Outer Space*和*Troll2*就因为这个原因而臭名昭著的。

简而言之，电影评分决不是固定不变的。 因此，使用时间动力学可以得到更准确的电影推荐 :cite: Koren.2009 。 当然，序列数据不仅仅是关于电影评分的。 下面给出了更多的场景：

- 在使用应用程序时，许多用户都有很强的特定习惯。 例如，在学生放学后社交媒体应用更受欢迎。在市场开放时股市交易软件更常用。
- 预测明天的股价要比过去的股价更困难，尽管两者都只是估计一个数字。 毕竟，先见之明比事后诸葛亮难得多。在统计学中，前者（对超出已知观测范围进行预测）称为*外推法*（extrapolation）， 而后者（在现有观测值之间进行估计）称为*内插法*（interpolation）。
- 在本质上，音乐、语音、文本和视频都是连续的。 如果它们的序列被我们重排，那么就会失去原有的意义。 比如，一个文本标题“狗咬人”远没有“人咬狗”那么令人惊讶，尽管组成两句话的字完全相同。
- 地震具有很强的相关性，即大地震发生后，很可能会有几次小余震， 这些余震的强度比非大地震后的余震要大得多。 事实上，地震是时空相关的，即余震通常发生在很短的时间跨度和很近的距离内。
- 人类之间的互动也是连续的，这可以从微博上的争吵和辩论中看出。

# 统计工具

处理序列数据需要统计工具和新的深度神经网络架构。 为了简单起见，以下图所示的股票价格（富时100指数）为例。



其中，用$x_t$表示价格，即在*时间步*（time step） $t \in \mathbb{Z}^+$时，观察到的价格$x_t$。 请注意，$t$对于本文中的序列通常是离散的，并在整数或其子集上变化。 假设一个交易员想在$t$日的股市中表现良好，于是通过以下途径预测$x_t$：

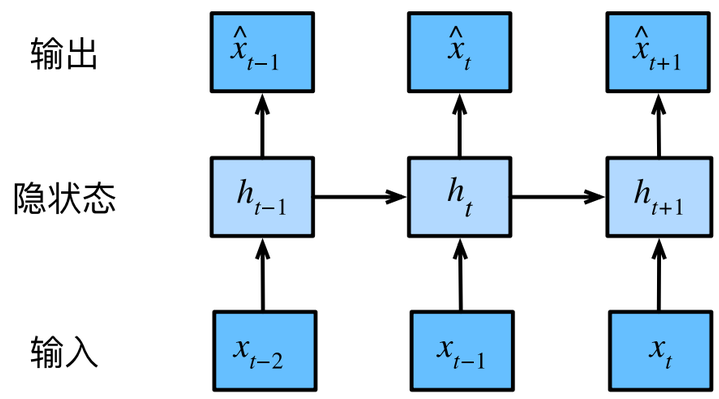
$$x_t \sim P(x_t \mid x_{t-1}, \ldots, x_1).$$

## 自回归模型

为了实现这个预测，交易员可以使用回归模型， 例如在线性回归中训练的模型。 仅有一个主要问题：输入数据的数量， 输入$x_{t-1}, \ldots, x_1$本身因$t$而异。 也就是说，输入数据的数量这个数字将会随着我们遇到的数据量的增加而增加，因此需要一个近似方法来使这个计算变得容易处理。 后面的大部分内容将围绕着如何有效估计$P(x_t \mid x_{t-1}, \ldots, x_1)$展开。 简单地说，它归结为以下两种策略。

第一种策略，假设在现实情况下相当长的序列 $x_{t-1}, \ldots, x_1$可能是不必要的， 因此我们只需要满足某个长度为$\tau$的时间跨度， 即使用观测序列$x_{t-1}, \ldots, x_{t-\tau}$。 当下获得的最直接的好处就是参数的数量总是不变的， 至少在$t > \tau$时如此，这就使我们能够训练一个上面提及的深度网络。 这种模型被称为*自回归模型*（autoregressive models）， 因为它们是对自己执行回归。

第二种策略，如下图所示， 是保留一些对过去观测的总结$h_t$， 并且同时更新预测$\hat{x}_t$和总结$h_t$。 这就产生了基于$\hat{x}_t = P(x_t \mid h_t)$估计$x_t$， 以及公式$h_t = g(h_{t-1}, x_{t-1})$更新的模型。 由于$h_t$从未被观测到，这类模型也被称为 *隐变量自回归模型*（latent autoregressive models）。




这两种情况都有一个显而易见的问题：如何生成训练数据？ 一个经典方法是使用历史观测来预测下一个未来观测。 显然，并不能指望时间会停滞不前。 然而，一个常见的假设是虽然特定值$x_t$可能会改变， 但是序列本身的动力学不会改变。 这样的假设是合理的，因为新的动力学一定受新的数据影响， 而不可能用目前所掌握的数据来预测新的动力学。 统计学家称不变的动力学为*静止的*（stationary）。 因此，整个序列的估计值都将通过以下的方式获得：

$$P(x_1, \ldots, x_T) = \prod_{t=1}^{T} P(x_t \mid x_{t-1}, \ldots, x_1).$$

注意，如果处理的是离散的对象（如单词）， 而不是连续的数字，则上述的考虑仍然有效。 唯一的差别是，对于离散的对象， 我们需要使用分类器而不是回归模型来估计$P(x_t \mid x_{t-1}, \ldots, x_1)$。

## 马尔可夫模型

回想一下，在自回归模型的近似法中， 使用$x_{t-1}, \ldots, x_{t-\tau}$ 而不是$x_{t-1}, \ldots, x_1$来估计$x_t$。 只要这种是近似精确的，就说序列满足*马尔可夫条件*（Markov condition）。 特别是，如果$\tau = 1$，得到一个 *一阶马尔可夫模型*（first-order Markov model） ， $P(x)$由下式给出：

$$P(x_1, \ldots, x_T) = \prod_{t=1}^{T} P(x_t \mid x_{t-1}) \text{ 当 } P(x_1 \mid x_0) = P(x_1).$$

当假设$x_t$仅是离散值时，这样的模型特别棒， 因为在这种情况下，使用动态规划可以沿着马尔可夫链精确地计算结果。 例如，可以高效地计算$P(x_{t+1} \mid x_{t-1})$：

$$\begin{aligned} P(x_{t+1} \mid x_{t-1}) &= \frac{\sum_{x_t} P(x_{t+1}, x_t, x_{t-1})}{P(x_{t-1})} \\ &= \frac{\sum_{x_t} P(x_{t+1} \mid x_t, x_{t-1}) P(x_t, x_{t-1})}{P(x_{t-1})} \\ &= \sum_{x_t} P(x_{t+1} \mid x_t) P(x_t \mid x_{t-1}) \end{aligned}$$

利用这一事实，只需要考虑过去观察中的一个非常短的历史： $P(x_{t+1} \mid x_t, x_{t-1}) = P(x_{t+1} \mid x_t)$。

## 因果关系

原则上，将$P(x_1, \ldots, x_T)$倒序展开也没什么问题。 毕竟，基于条件概率公式，我们总是可以写出：

$$P(x_1, \ldots, x_T) = \prod_{t=T}^{1} P(x_t \mid x_{t+1}, \ldots, x_T).$$

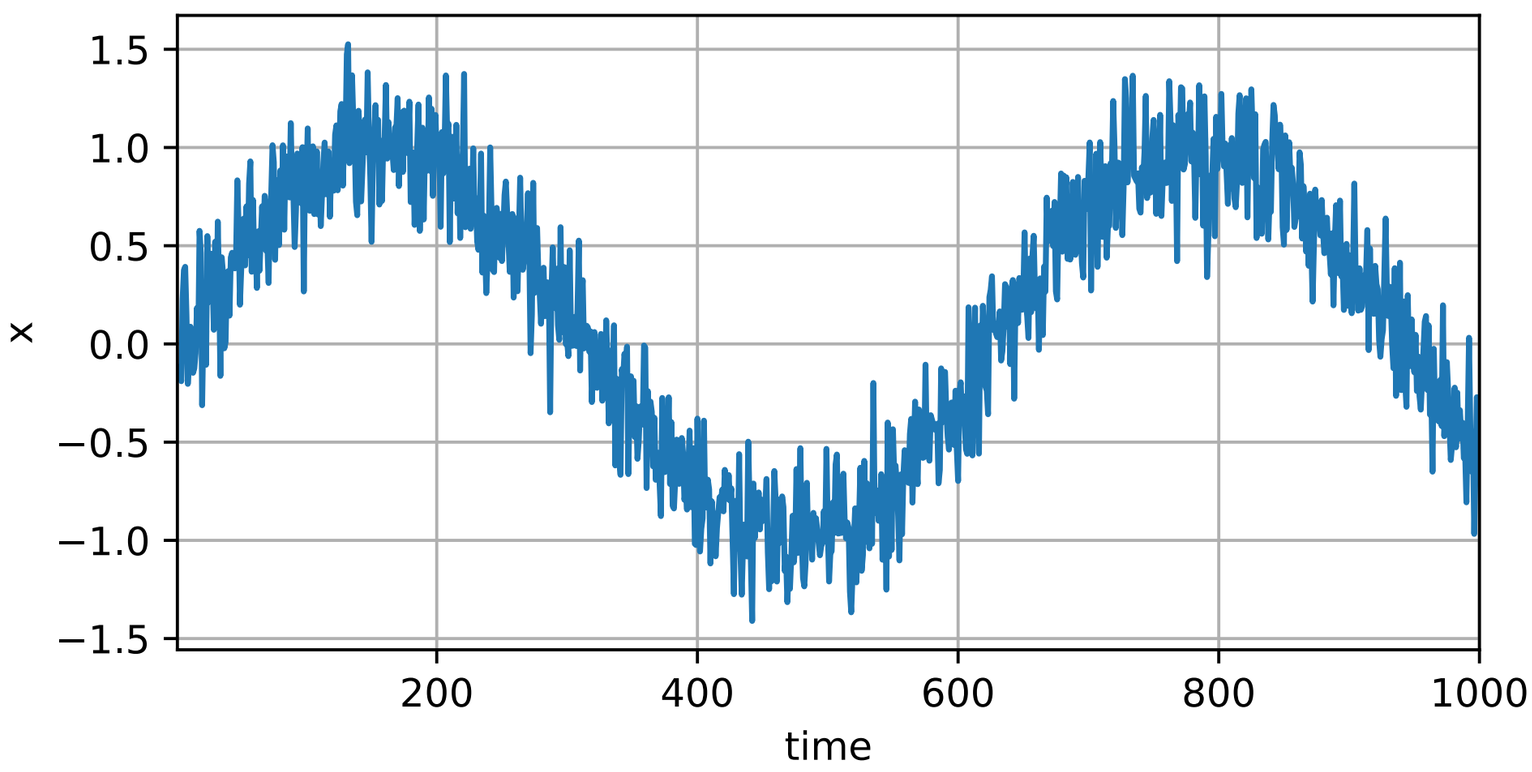
事实上，如果基于一个马尔可夫模型， 还可以得到一个反向的条件概率分布。 然而，在许多情况下，数据存在一个自然的方向，即在时间上是前进的。 很明显，未来的事件不能影响过去。 因此，如果改变$x_t$， 可能会影响未来发生的事情$x_{t+1}$，但不能反过来。 也就是说，如果改变$x_t$，基于过去事件得到的分布不会改变。 因此，解释$P(x_{t+1} \mid x_t)$应该比解释$P(x_t \mid x_{t+1})$更容易。 例如，在某些情况下，对于某些可加性噪声$\epsilon$， 显然可以找到$x_{t+1} = f(x_t) + \epsilon$， 而反之则不行。 这是个好消息，因为这个前进方向通常也是我们感兴趣的方向。

## 训练

在了解了上述统计工具后，在实践中尝试一下！ 首先，生成一些数据：(**使用正弦函数和一些可加性噪声来生成序列数据， 时间步为**$1, 2, \ldots, 1000$。)

In [1]:
```
%matplotlib inline
import torch
from torch import nn
from d2l import torch as d2l
```

In [2]:
```
T = 1000  # 总共产生1000个点
time = torch.arange(1, T + 1, dtype=torch.float32)
x = torch.sin(0.01 * time) + torch.normal(0, 0.2, (T,))
d2l.plot(time, [x], 'time', 'x', xlim=[1, 1000], figsize=(6, 3))
```




接下来，将这个序列转换为模型的“特征 - 标签”（feature-label）对。 基于嵌入维度$\tau$，**将数据映射为数据对**$y_t = x_t$ **和**$\mathbf{x}_t = [x_{t-\tau}, \ldots, x_{t-1}]$。

这比提供的数据样本少了$\tau$个， 因为没有足够的历史记录来描述前$\tau$个数据样本。 一个简单的解决办法是：如果拥有足够长的序列就丢弃这几项； 另一个方法是用零填充序列。 在这里，仅使用前600个“特征 - 标签”对进行训练。

In [3]:
```
tau = 4
features = torch.zeros((T - tau, tau))
for i in range(tau):
    features[:, i] = x[i: T - tau + i]
labels = x[tau:].reshape((-1, 1))
```

In [4]:
```
batch_size, n_train = 16, 600
# 只有前n_train个样本用于训练
train_iter = d2l.load_array((features[:n_train], labels[:n_train]),
                            batch_size, is_train=True)
```

在这里，**使用一个相当简单的架构训练模型： 一个拥有两个全连接层的多层感知机**，ReLU激活函数和平方损失。

In [5]:
```
# 初始化网络权重的函数
def init_weights(m):
    if type(m) == nn.Linear:
        nn.init.xavier_uniform_(m.weight)

# 一个简单的多层感知机
def get_net():
    net = nn.Sequential(nn.Linear(4, 10),
                        nn.ReLU(),
                        nn.Linear(10, 1))
    net.apply(init_weights)
    return net

# 平方损失。注意：MSELoss计算平方误差时不带系数1/2
loss = nn.MSELoss(reduction='none')
```

现在，准备**训练模型**了。实现下面的训练代码的方式与之前训练基本相同。

In [6]:
```
def train(net, train_iter, loss, epochs, lr):
    trainer = torch.optim.Adam(net.parameters(), lr)
    for epoch in range(epochs):
        for X, y in train_iter:
            trainer.zero_grad()
            l = loss(net(X), y)
            l.sum().backward()
            trainer.step()
        print(f'epoch {epoch + 1}, '
              f'loss: {d2l.evaluate_loss(net, train_iter, loss):f}')

net = get_net()
train(net, train_iter, loss, 5, 0.01)
```

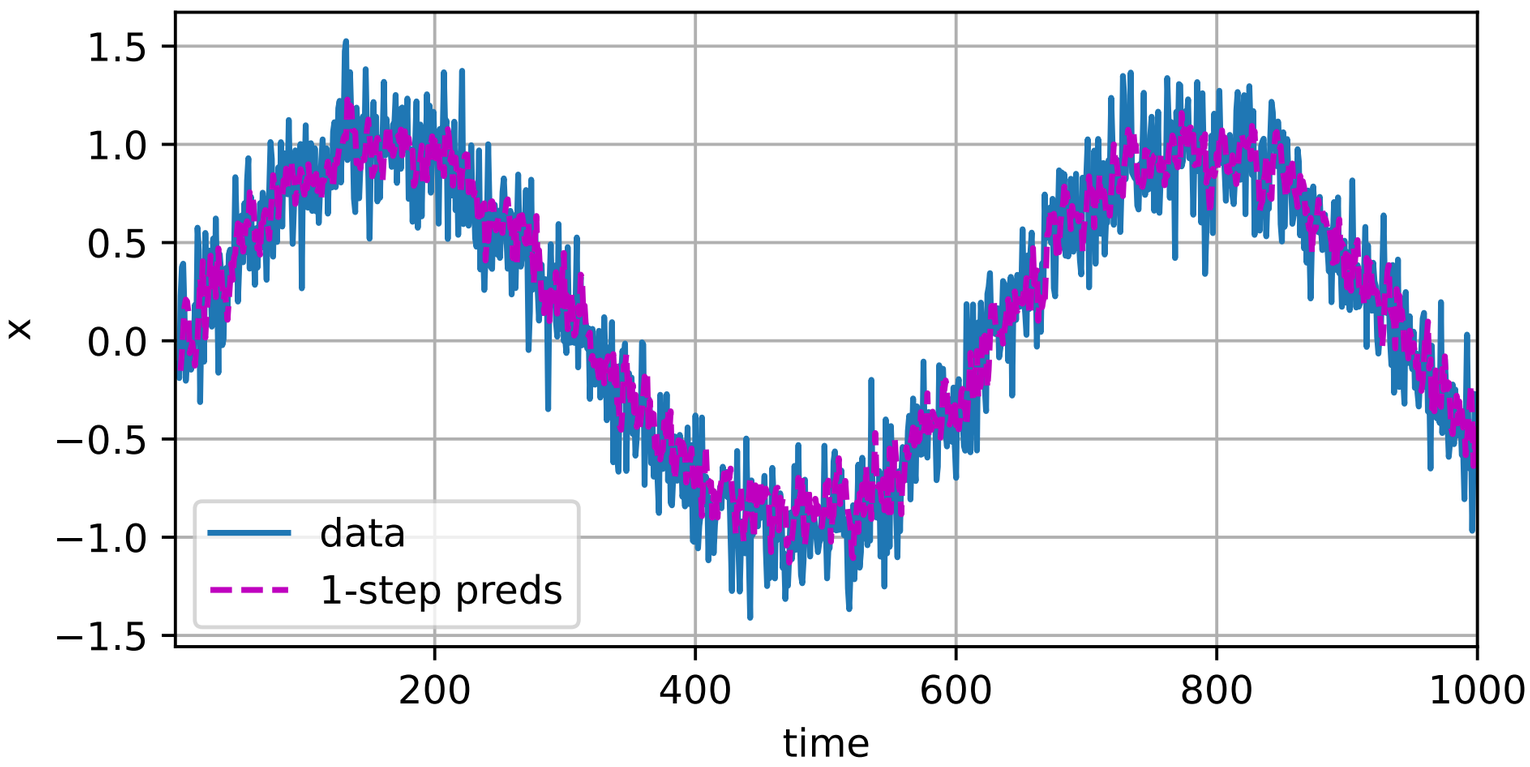
```
epoch 1, loss: 0.060049
epoch 2, loss: 0.055132
epoch 3, loss: 0.054419
epoch 4, loss: 0.051754
epoch 5, loss: 0.052902
```

## 预测

训练损失很小，希望模型能有很好的工作效果。

看看这在实践中意味着什么。 首先是检查**模型预测下一个时间步**的能力， 也就是*单步预测*（one-step-ahead prediction）。

In [7]:
```
onestep_preds = net(features)
d2l.plot([time, time[tau:]],
         [x.detach().numpy(), onestep_preds.detach().numpy()], 'time',
         'x', legend=['data', '1-step preds'], xlim=[1, 1000],
         figsize=(6, 3))
```



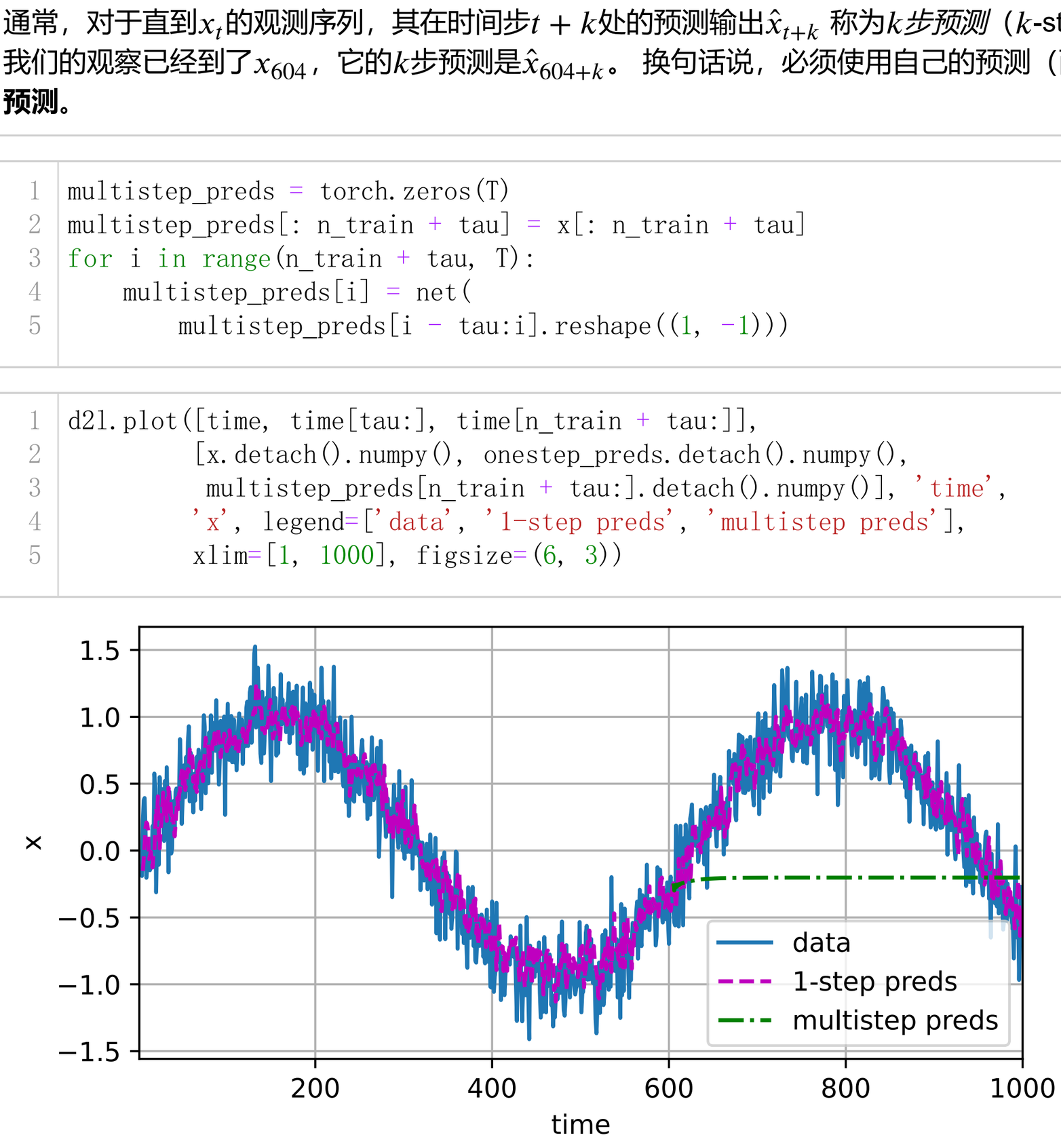

正如我们所料，单步预测效果不错。 即使这些预测的时间步超过了$600+4$（`n_train + tau`），其结果看起来仍然是可信的。 然而有一个小问题：如果数据观察序列的时间步只到$604$，我们需要一步一步地向前迈进：

$$\begin{aligned}\hat{x}_{605} &= f(x_{601}, x_{602}, x_{603}, x_{604}),\\ \hat{x}_{606} &= f(x_{602}, x_{603}, x_{604}, \hat{x}_{605}),\\ \hat{x}_{607} &= f(x_{603}, x_{604}, \hat{x}_{605}, \hat{x}_{606}),\\ \hat{x}_{608} &= f(x_{604}, \hat{x}_{605}, \hat{x}_{606}, \hat{x}_{607}),\\ \hat{x}_{609} &= f(\hat{x}_{605}, \hat{x}_{606}, \hat{x}_{607}, \hat{x}_{608}),\\ &\ldots\end{aligned}$$

通常，对于直到$x_t$的观测序列，其在时间步$t+k$处的预测输出$\hat{x}_{t+k}$称为*$k$步预测*（$k$-step-ahead-prediction）。 由于我们的观察已经到了$x_{604}$，它的$k$步预测是$\hat{x}_{604+k}$。 换句话说，必须使用自己的预测（而不是原始数据）来**进行多步预测。**

In [8]:
```
multistep_preds = torch.zeros(T)
multistep_preds[: n_train + tau] = x[: n_train + tau]
for i in range(n_train + tau, T):
    multistep_preds[i] = net(
        multistep_preds[i - tau:i].reshape((1, -1)))
```

In [9]:
```
d2l.plot([time, time[tau:], time[n_train + tau:]],
         [x.detach().numpy(), onestep_preds.detach().numpy(),
          multistep_preds[n_train + tau:].detach().numpy()], 'time',
         'x', legend=['data', '1-step preds', 'multistep preds'],
         xlim=[1, 1000], figsize=(6, 3))
```



如上面的例子所示，绿线的预测显然并不理想。 经过几个预测步骤之后，预测的结果很快就会衰减到一个常数。 为什么这个算法效果这么差呢?

事实是由于错误的累积： 假设在步骤1之后，积累了一些错误$\epsilon_1 = \bar{\epsilon}$。 于是，步骤2的输入被扰动了$\epsilon_1$， 结果积累的误差是依照次序的$\epsilon_2 = \bar{\epsilon} + c\epsilon_1$， 其中$c$为某个常数，后面的预测误差依此类推。 因此误差可能会相当快地偏离真实的观测结果。 例如，未来24小时的天气预报往往相当准确， 但超过这一点，精度就会迅速下降。

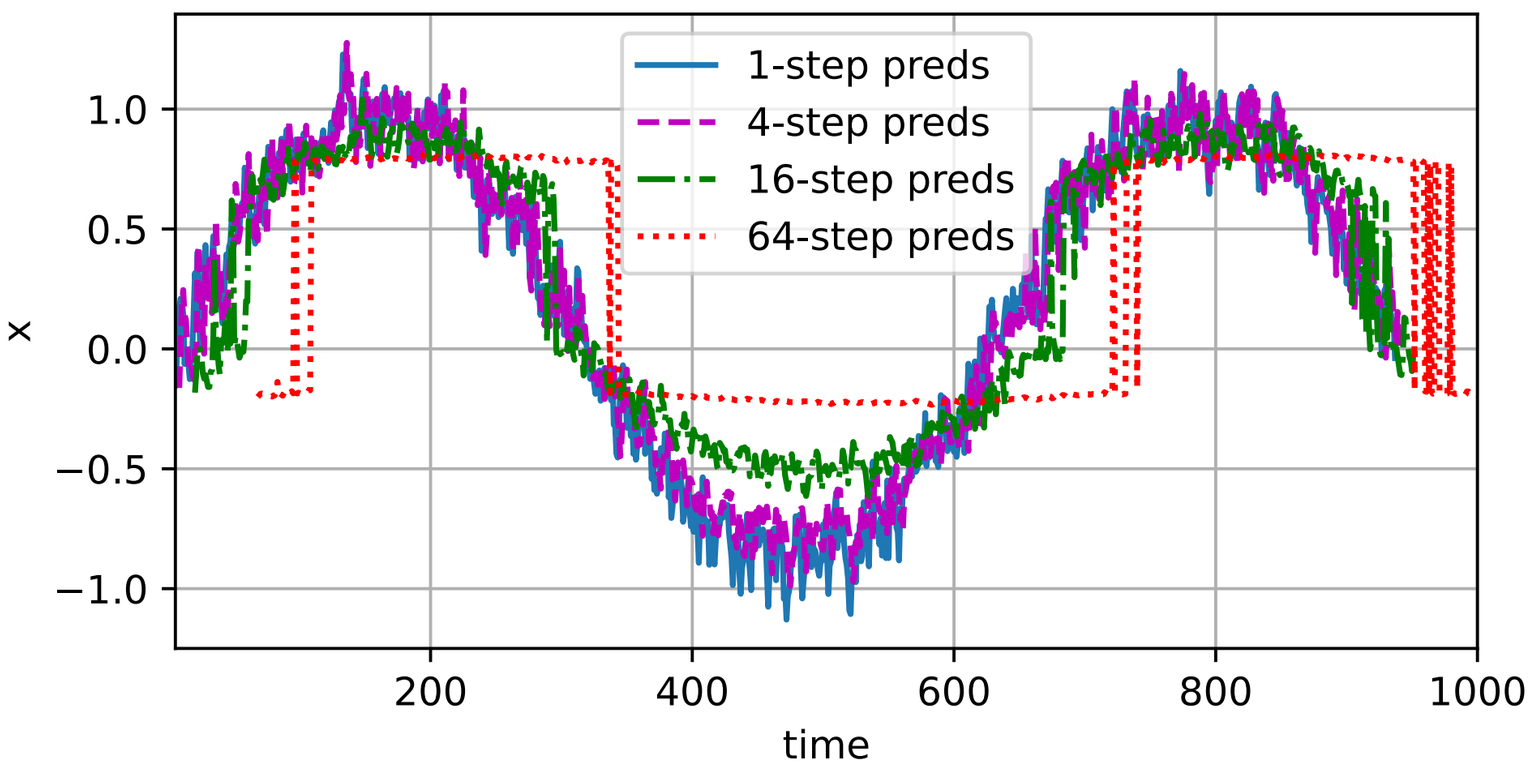
基于$k = 1, 4, 16, 64$，通过对整个序列预测的计算， 让我们**更仔细地看一下$k$步预测**的困难。

In [10]:
```
max_steps = 64
```

In [11]:
```
features = torch.zeros((T - tau - max_steps + 1, tau + max_steps))
# 列i（i<tau）是来自x的观测，其时间步从（i+1）到（i+T-tau-max_steps+1）
for i in range(tau):
    features[:, i] = x[i: i + T - tau - max_steps + 1]

# 列i（i>=tau）是来自（i-tau+1）步的预测，其时间步从（i+1）到（i+T-tau-max_steps+1）
for i in range(tau, tau + max_steps):
    features[:, i] = net(features[:, i - tau:i]).reshape(-1)
```

In [12]:
```
steps = (1, 4, 16, 64)
d2l.plot([time[tau + i - 1: T - max_steps + i] for i in steps],
         [features[:, (tau + i - 1)].detach().numpy() for i in steps], 'time', 'x',
         legend=[f'{i}-step preds' for i in steps], xlim=[5, 1000],
         figsize=(6, 3))
```




以上例子清楚地说明了当试图预测更远的未来时，预测的质量是如何变化的。 虽然“4步预测”看起来仍然不错，但超过这个跨度的任何预测几乎都是无用的。