Intro

Paradigms

- Supervised Learning

  $Given\ D = \{X_i, Y_i\},\ learn\ f(\cdot) : Y_i = f(X_i),\ s.t. D^{new} = \{X_j\} => \{Y_j\}$
- Unsupervised Learning

  $Given\ D = \{X_i\},\ learn\ f(\cdot) : Y_i = f(X_i),\ s.t. D^{new} = \{X_j\} => \{Y_j\}$

Example

Polynomial curve fitting

Fit the data using a polynomial function of the form:

$$y(x, \mathbf{w}) = \mathbf{w_0 + w_1 x + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j}$$

$\mathbf{w}$ is the parameters we need to adapt according to dataset $\{(x_n, y_n)\}_N$

Minimize "loss function" to find the w:

$$\mathbf{w = argmin_w \{E(w)\}, E(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2}$$

Overfitting

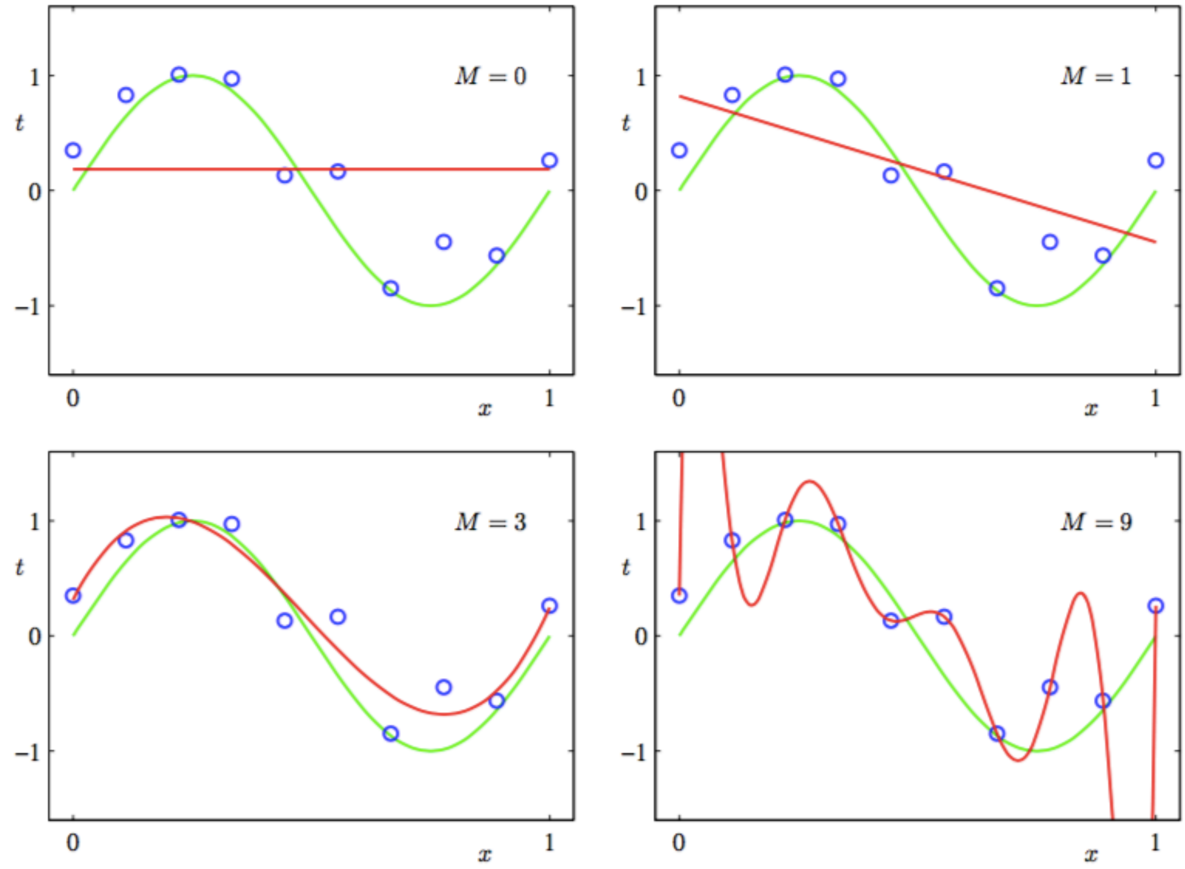

Figure 1.4 Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

For M = 9, the training set error goes to zero, while test set error become very large due to overfitting. The reason is that we have 10 coefficients(w_0 to w_9) thus containing **10 degrees of freedom, and so they can be tuned exactly to the 10 data points in the training set**.

Avoid overfitting(1)

More data

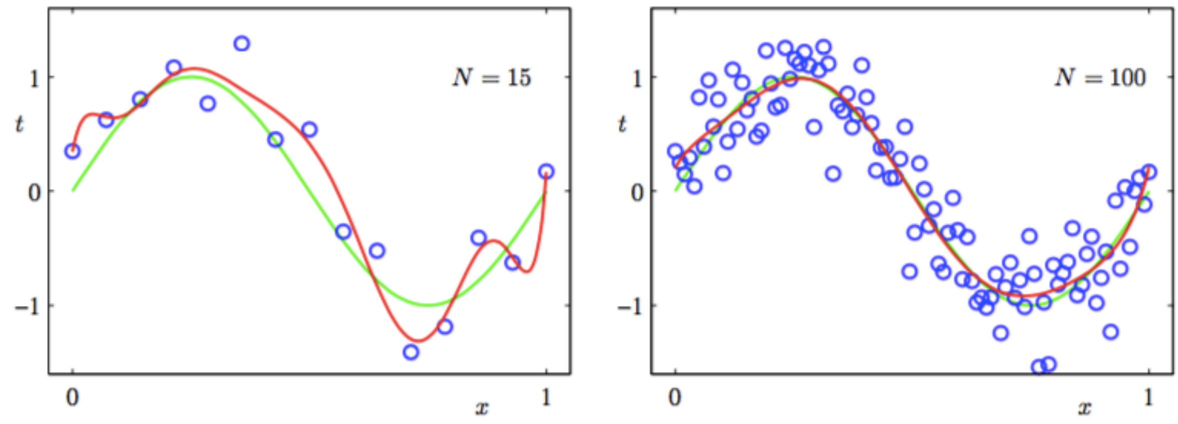

Figure 1.6 Plots of the solutions obtained by minimizing the sum-of-squares error function using the $M = 9$ polynomial for $N = 15$ data points (left plot) and $N = 100$ data points (right plot). We see that increasing the size of the data set reduces the over-fitting problem.

Avoid overfitting(2)

Loss function with **panalty item(or regularization)** on ||w||

$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - \mathbf{t_n}\}^2 + \frac{\lambda}{2}||\mathbf{w}||^2$

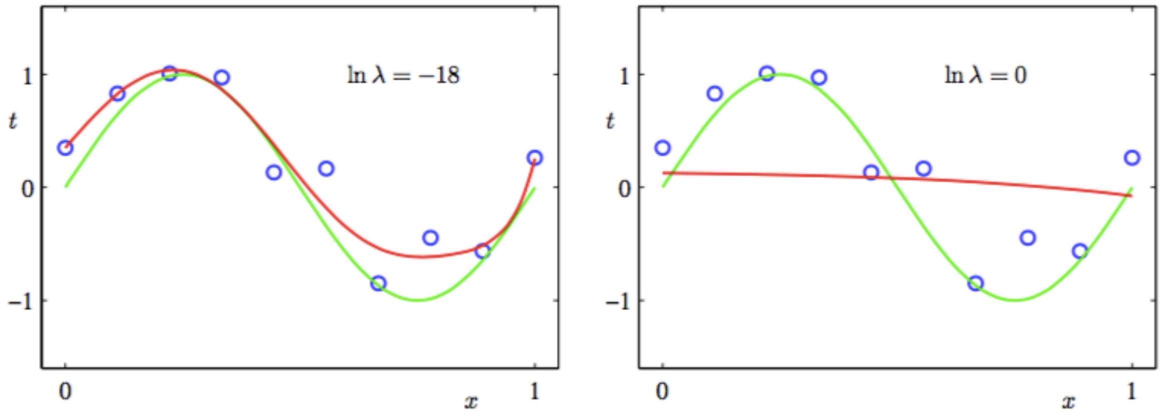

Figure 1.7 Plots of $M = 9$ polynomials fitted to the data set shown in Figure 1.2 using the regularized error function (1.4) for two values of the regularization parameter λ corresponding to $\ln\lambda = -18$ and $\ln\lambda = 0$. The case of no regularizer, i.e., $\lambda = 0$, corresponding to $\ln\lambda = -\infty$, is shown at the bottom right of Figure 1.4.

Probability Theory

Rules of Probability

- **sum rule:** $p(Y) = \sum_{Y} p(X, Y)$
- **product rule:** $p(Y, X) = p(Y|X)P(X)$
- **Bayes' theorem:** $p(Y|X) = \frac{p(X|Y)p(Y)}{P(X)},\ P(X) = \sum_{Y} p(X|Y)p(Y)$

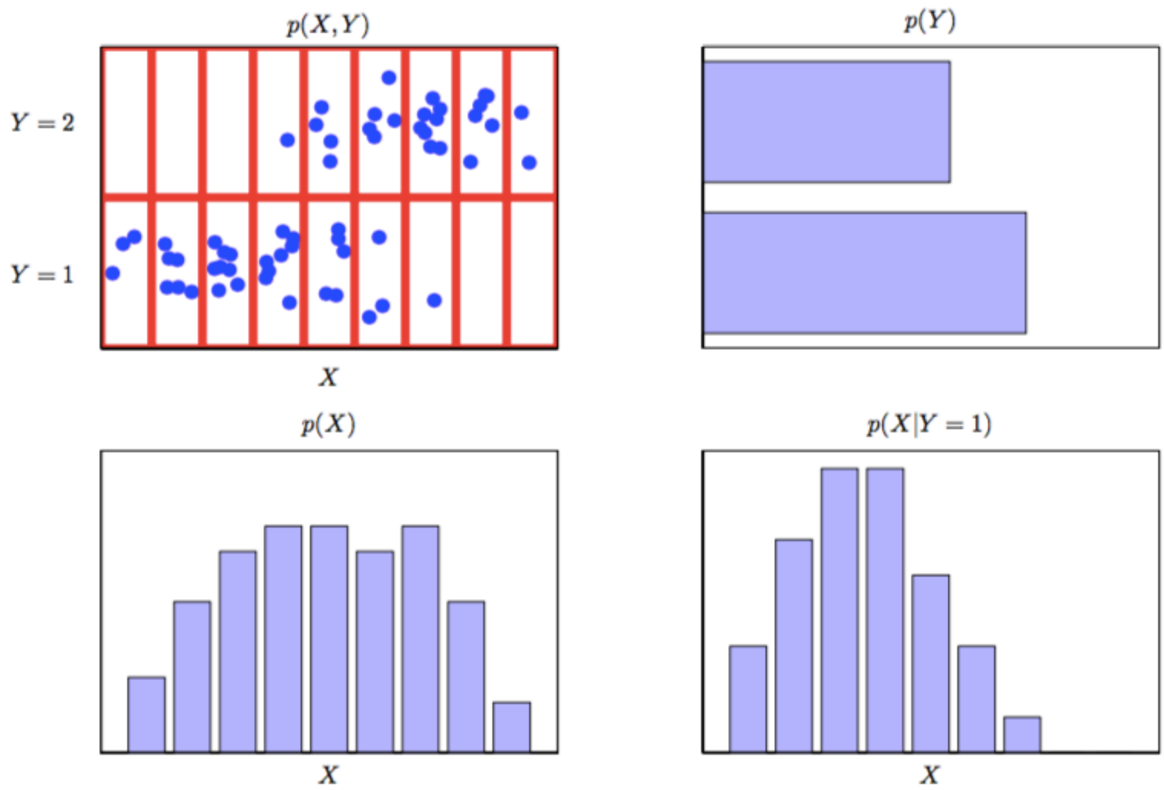

Figure 1.11 An illustration of a distribution over two variables, X, which takes 9 possible values, and Y, which takes two possible values. The top left figure shows a sample of 60 points drawn from a joint probability distribution over these variables. The remaining figures show histogram estimates of the marginal distributions $p(X)$ and $p(Y)$, as well as the conditional distribution $p(X|Y = 1)$ corresponding to the bottom row in the top left figure.

Probability densities

$p(x \in (a,b)) = \int_a^b p(x)dx,\ p(x) \geq 0:\ density\ function$

- Note: Under a **nonlinear change of variable**, a probability density transforms differently from a simple function, due to the **Jacobian factor**.

$$\begin{aligned} given\ x &= g(y) \\ \because p_x(x)dx &\simeq p_y(y)dy \\ \therefore p_y(y) &= p_x(x)|\frac{dx}{dy}| \\ &= p_x(g(y)) \end{aligned}$$

One consequence of this property is that the concept of the maximum of a probability density is **dependent on the choice of variable**.

Expectations and covariances

$E[f] = \sum_x p(x)f(x),\ E[f] = \int p(x)f(x)dx$

$var[f] = E[(f(x) - E[f])^2] = E[f^2] - E[f(x)]^2$

$cov[x,y] = E_{x,y}[(x - E[x])(y - E[y])] = E_{x,y}[xy] - E[x]E[y]$

$cov[\mathbf{x},\mathbf{y}] = \mathbf{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x} - \mathbf{E}[\mathbf{x}]\mathbf{y}^\mathbf{T} - \mathbf{E}[\mathbf{y}^\mathbf{T}]] = \mathbf{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^\mathbf{T}] - \mathbf{E}[\mathbf{x}]\mathbf{E}[\mathbf{y}^\mathbf{T}]$

Bayes' View

Bayes' theorem was used to **convert a prior probability into a posterior probability by incorporating the evidence provided by the observed data**.

Prior probability can be regarded as **knowledge gained before or "common sense"**.

From frequests' view, the **w** learned from dataset is fixed(by maximize likelihood function), while From Bayes' view, it's an uncertain variable represented by a probability distribution $p(\mathbf{w})$
.

Common path of Bayes' learning:

Loop

1. prior: $p(\mathbf{w})$
2. Observed dataset: $D = t_1, \ldots, t_N$
3. Posterior: $p(\mathbf{w}|\mathbf{D}) = \frac{\mathbf{p(D|w)p(w)}}{\mathbf{p(D)}}$ and regard it as new prior(updated by observations).

$posterior \propto likelihood \times prior$

$p(D) = \int \mathbf{p(D|w)p(w)dw}$

Gaussian distribution

$\mathcal{N}(x|u, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2\sigma^2}(x-u)^2\}$

$\mathcal{N}(x|u, \sum) = \frac{1}{\sqrt{(2\pi)^D|\sum|}} exp\{-\frac{1}{2}(x-u)^T \sum(x-u)\}$