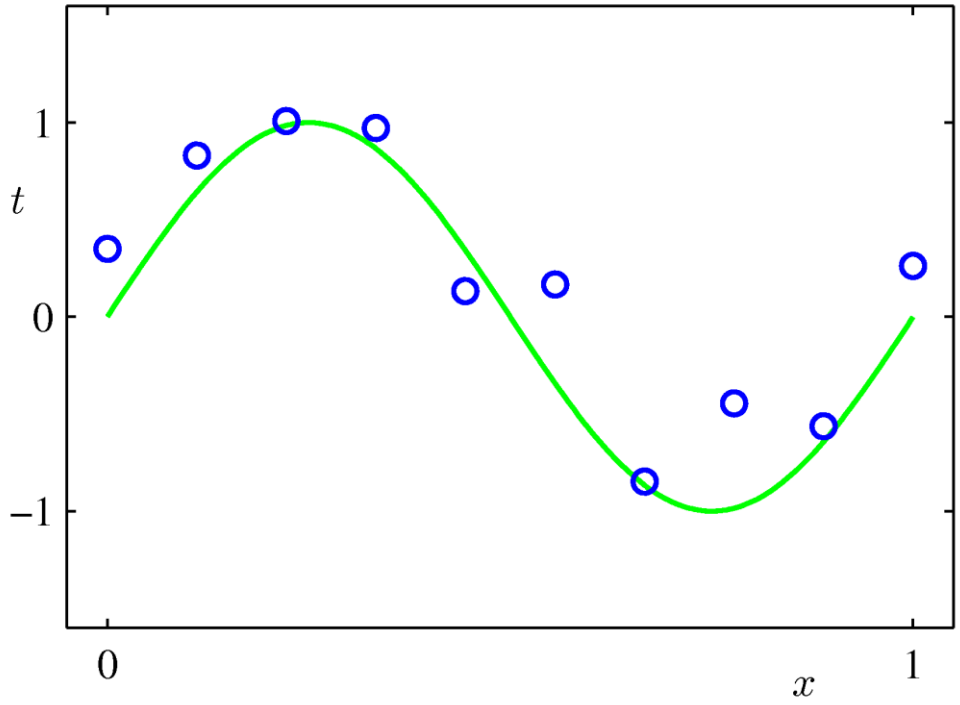
パターン認識と機械学習
第1章：序論（前半）

Christopher M. Bishop (2006):
Pattern Recognition and Machine Learning, Springer, pp.1-37

1. 例：多項式曲線フィッティング



問：図の青点(訓練集合)*にうまくフィットする曲線はどのような式になるか？

*緑線（$\sin(2\pi x)$）から正規分布に従うランダムノイズ（誤差や観測されない信号元の変動にあたる）を加えて生成したもの

とりあえず、以下のようなM次の多項式の係数を求めることでフィッティングを行うことにする。

$$y(x,\mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

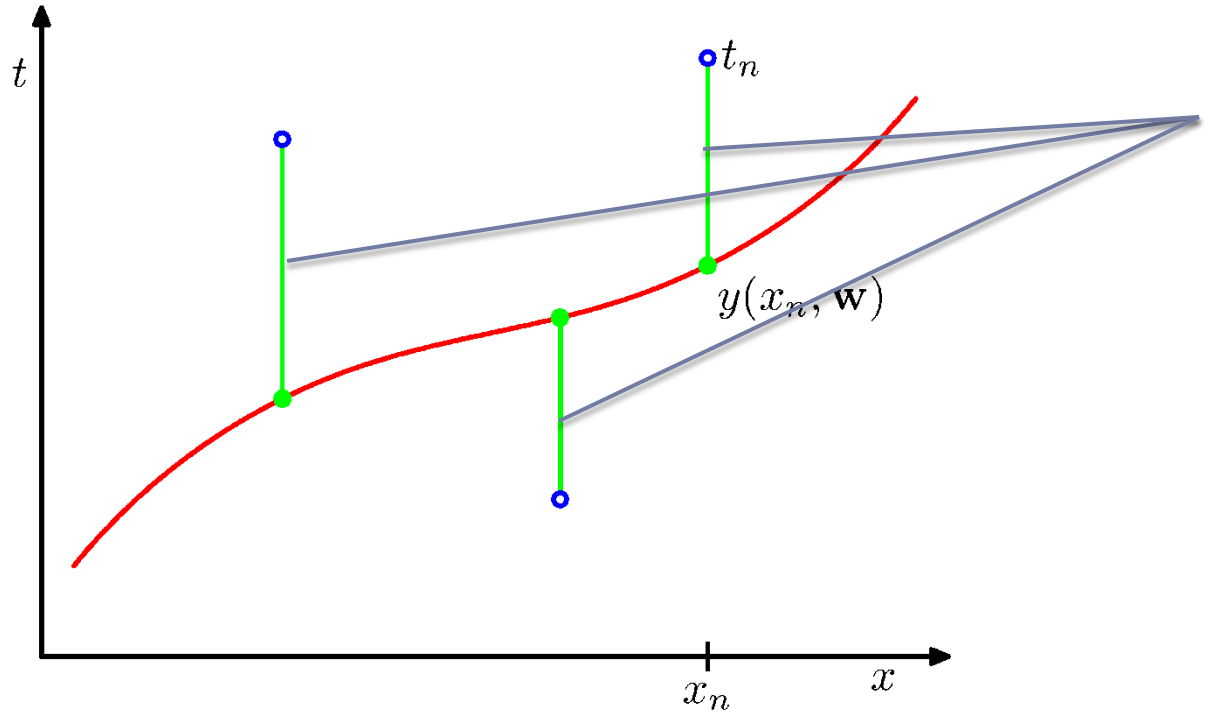
※上式は
・多項式 $y(x,\mathbf{w})$ はx の非線形関数であるが…
・wの線形関数である

…このように未知パラメータに関して線形であるような関数は線形モデルと呼ばれる。

では、最適なフィッティングを達成する係数をどのように求めるのか？

以下の誤差関数 $E(\mathbf{w})$、すなわち「各データ点 x_n における予測値 $y(x_n,\mathbf{w})$ 」と「対応する目標値 t_n 」との二乗誤差を最小化するような $\mathbf{w}$ を求める。

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



これらの和を最小化する！

この誤差関数は w についての2次関数だから解くのは簡単。
係数に関する微分は w にする1次式になり、この式を最小にするただ一つの解（以下では w* とする）を持つ。

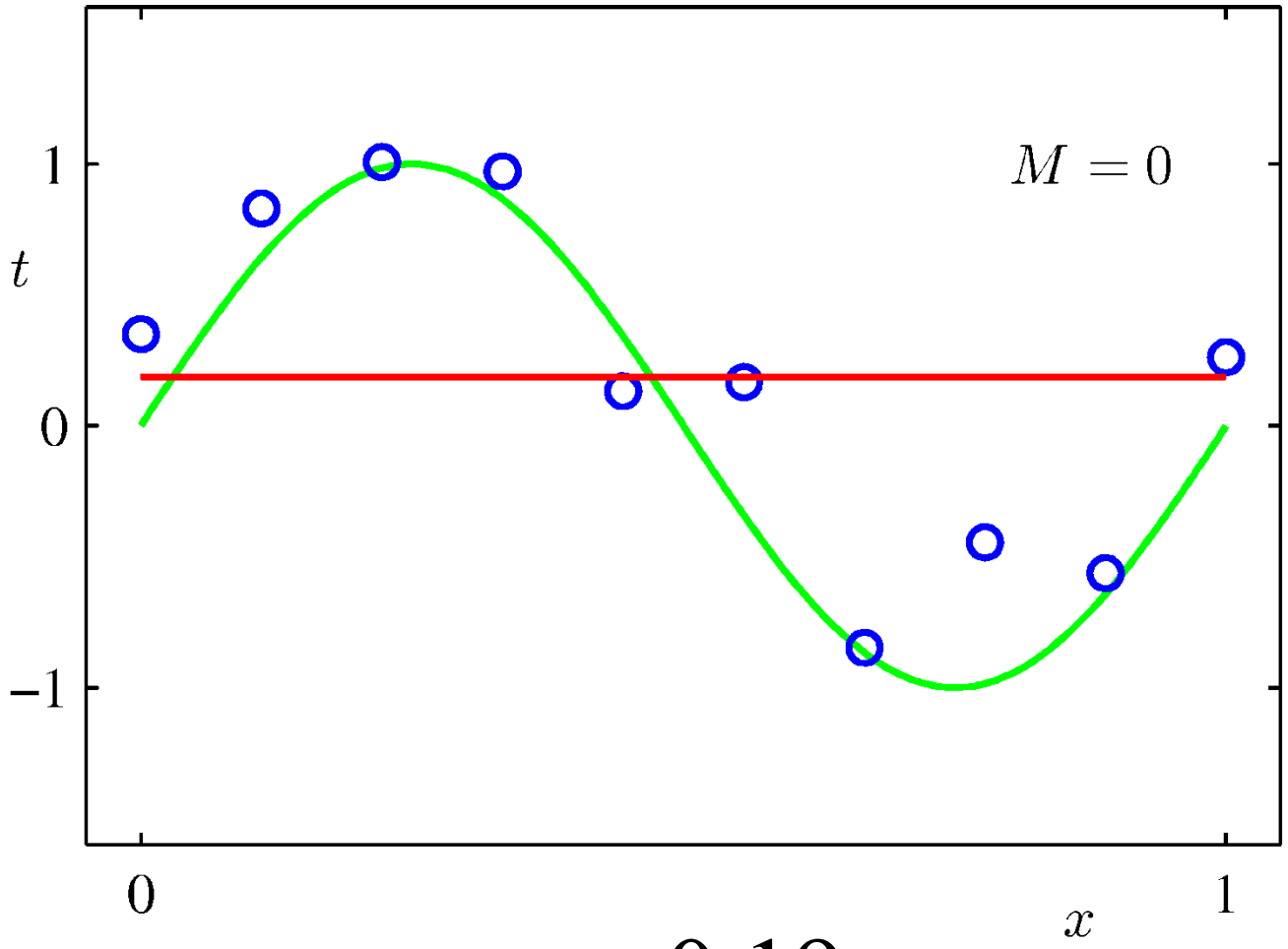
従って、求める多項式はこのw*を使って以下のように表される。

$$y(x, \mathbf{w}^*) = w_0 * + w_1 * x + w_2 * x^2 + \ldots + w_M * x^M = \sum_{j=0}^{M} w_j * x^j$$

残った問題は…式の次数*M*としてどの値を選ぶか？ということ。
すなわちデータ選択（データ比較）の問題。

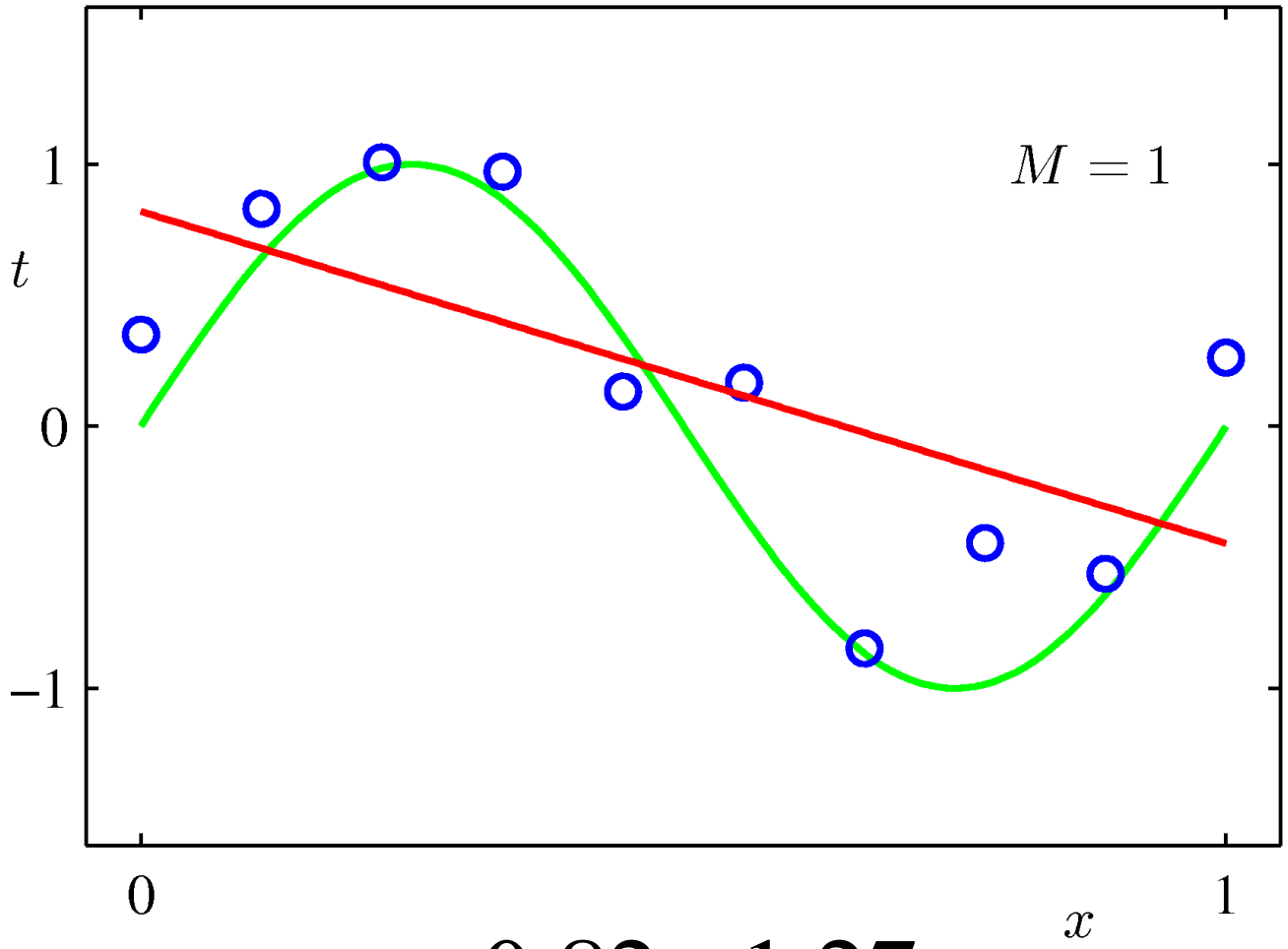
Mをいろいろ変えて見てみる。

M=0 の場合…定数なのでとうぜん全く合わない。



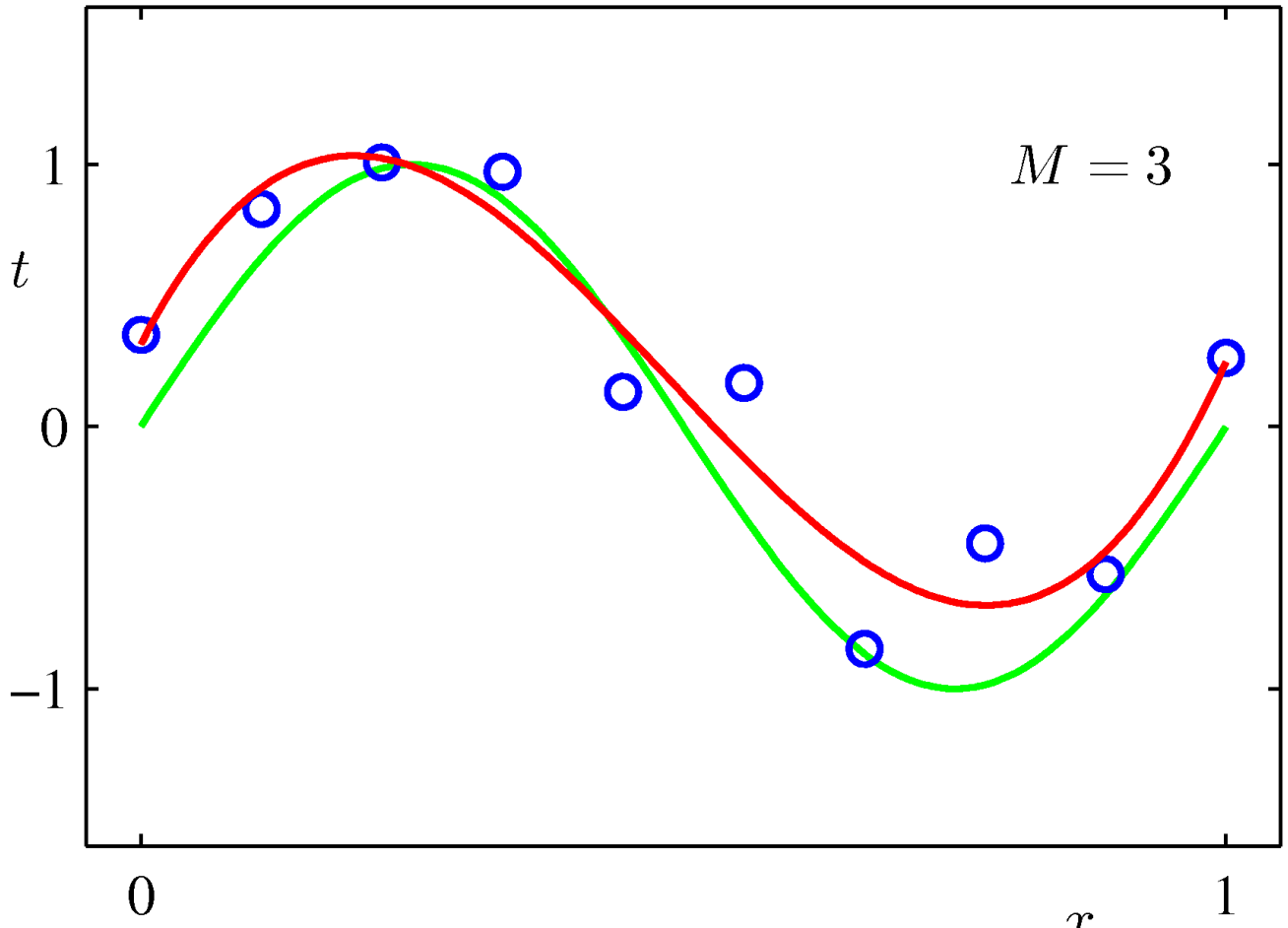
$$y = 0.19$$

M=1 の場合…1次関数なのでやっぱり合わない。



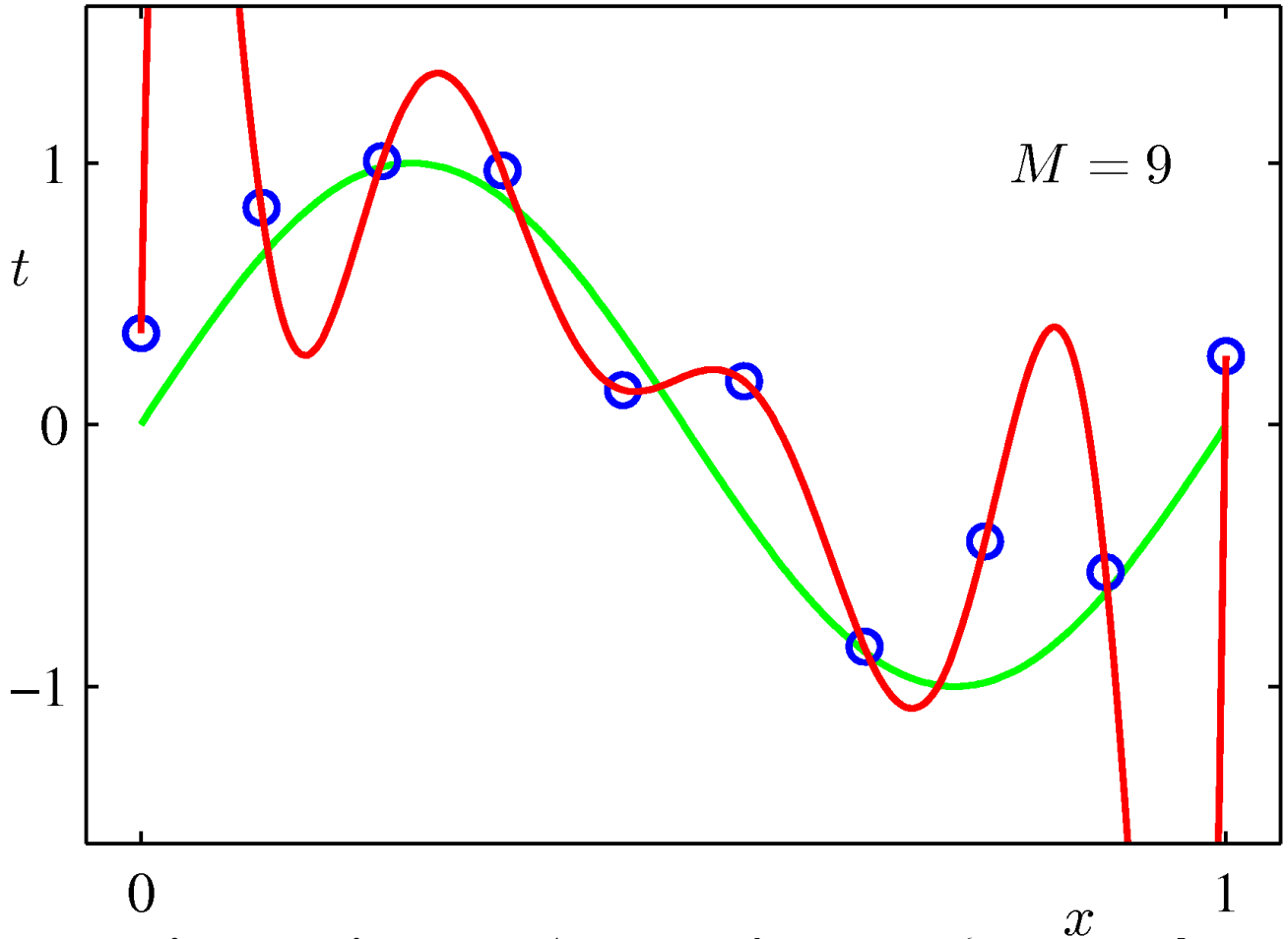
$$y = 0.82 - 1.27x$$

M=3 の場合…だいぶフィットしてきた！



$$y = 0.31 + 7.99x - 25.43x^2 + 17.37x^3$$

M=9 の場合… 誤差関数は0になったけれど、元の緑線に合っていない。
つまり…新たなテスト集合(緑線に近い分布をする)が与えられたときの予測の精度が低い…！



$$y = 0.35 + 232.37x - 5321.83x^2 + 48568.31x^3 - 231639.30x^4 + 640042.26x^5 - 1061800.18x^6 + 1042400.18x^7 - 557682.99x^8 + 125201.43x^9$$

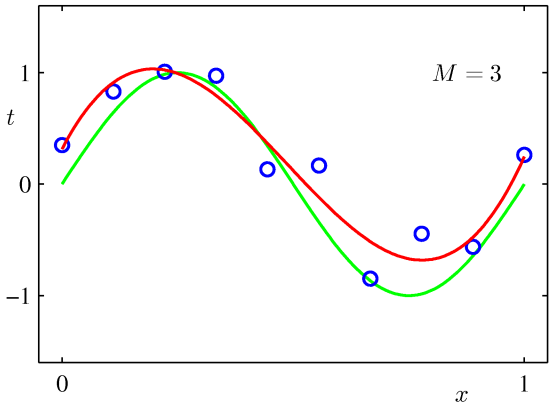
パラメータをたくさんとりすぎると
訓練集合のランダムノイズに引きずられてしまう。

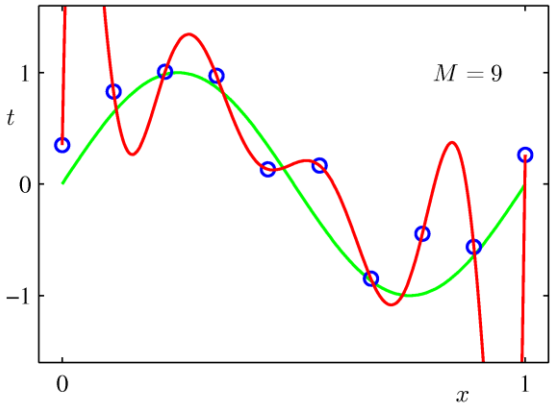
…これが過学習！

	M=3	M=9
訓練集合に対する誤差 (平均二乗平方根誤差*)	○	◎
テスト集合に対する誤差 (平均二乗平方根誤差*)	○	×

いかに青点から外れていないか

いかに緑線から外れていないか



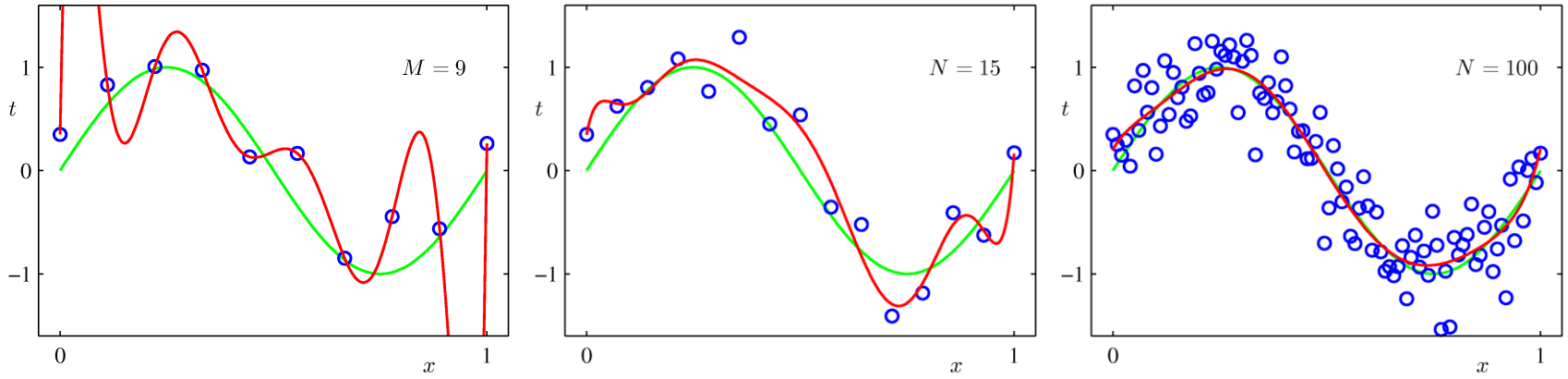


*平均二乗平方根誤差：以下の式で表され、違うMを用いた場合の誤差同士を比較することができる。

$$E_{RMS} = \sqrt{2E(\mathbf{w}^*)/N}$$

どうすれば過学習の問題を回避できるのか？

1.（パラメータ数に対する）訓練集合のサイズを増やす



問題点：モデルのパラメータ数は解くべき問題の複雑さに応じて選ぶのがもっともなはずだが、この方法をとると入手できる訓練集合のサイズに応じてモデルのパラメータ数を制限しなければならなくなる。

2. ベイズ的アプローチをとる

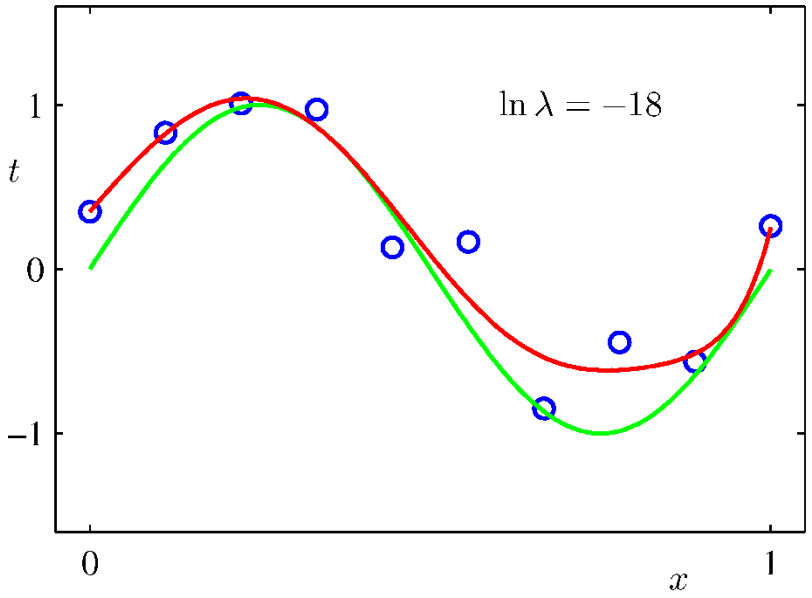
3. 正則化を行う

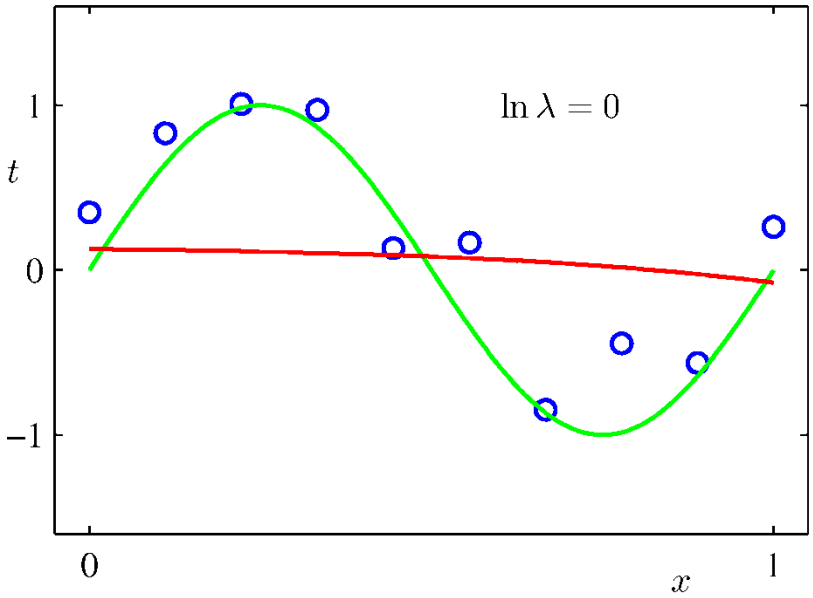
正則化とは誤差関数に罰金項を負荷し、係数が大きくなりすぎることを防ぐこと。
例えば、以下の右辺第2項が罰金項である。

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

ただしここで $\|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w} = w_0^2 + w_1^2 + \ldots + w_M^2$であり*、$\lambda$ は正則化項と二乗誤差の和の項の相対的な重要度を示す。
当然 λ の選び方によってあてはまりかたも変わってくる。

*係数 w_0 は目的関数の原点の選び方に依存しているため正則化から外すことも多い。





2. ベイズの定理とベイズ確率

まずはベイズの定理の復習から…
確率の乗法定理

$$p(X,Y) = p(Y \mid X)p(X)$$

*XかつYである確率=Xである確率×XのもとでのYの確率

より、

$$p(Y \mid X) = \frac{p(X,Y)}{p(X)}$$

であり、さらに

$$p(X,Y) = p(Y,X) = p(X \mid Y)p(Y)$$

であることから、以下のベイズの定理が成り立つ

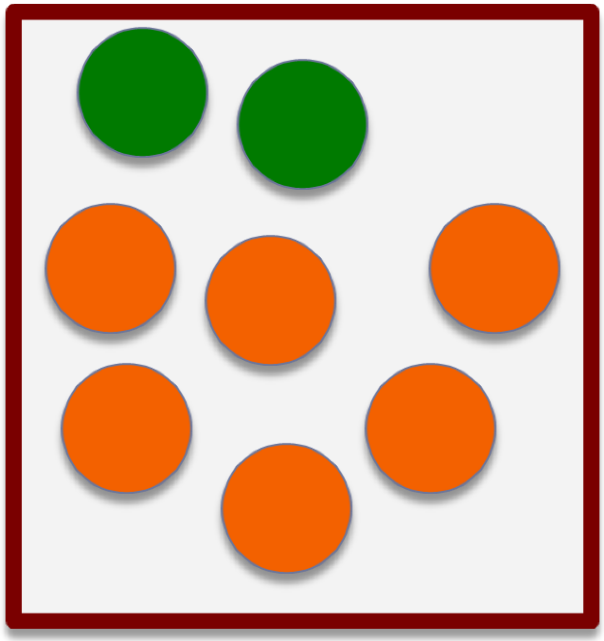
$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)}$$

もうちょっとベイズの定理に慣れましょう。
例:
赤と青のいずれかの箱が置いてあり(観測者にはそれがどちらであるかは分からない)その箱の中から果物を取り出す。

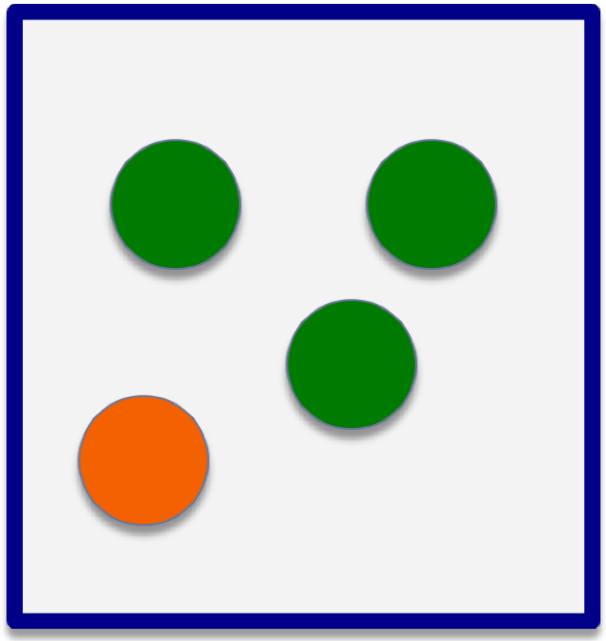
また、以下のことが分かっている。

赤い箱である確率は2/5
赤い箱にはりんごが2個とオレンジが6個

青い箱である確率は3/5
青い箱にはりんごが3個とオレンジが1個



2/5



3/5

ここで、箱から果物を取り出してみると、オレンジであったとする。
このとき置いてある箱が赤い箱である確率はどのくらいだろうか？

事前に分っている情報を整理すると…

$p(B=r)=2/5$ …赤い箱を選ぶ確率
$p(B=b)=3/5$ …青い箱を選ぶ確率
$p(F=a\,|\,B=r)=2/8$ …赤い箱からりんごを選ぶ確率
$p(F=o\,|\,B=r)=6/8$ …赤い箱からオレンジを選ぶ確率
$p(F=a\,|\,B=b)=3/4$ …青い箱からりんごを選ぶ確率
$p(F=o\,|\,B=b)=1/4$ …青い箱からオレンジを選ぶ確率

また、ここから次の確率も容易に求めることができる*

$$p(F=a)=p(F=a\,|\,B=r)p(B=r)+p(F=a\,|\,B=b)p(B=b)=11/20$$
$$p(F=o)=p(F=o\,|\,B=r)p(B=r)+p(F=o\,|\,B=b)p(B=b)=9/20$$

*たとえば前者の式は…
りんごを選ぶ確率=(赤い箱からりんごを選ぶ確率)*(赤い箱を選ぶ確率)
+(青い箱からりんごを選ぶ確率)*(青い箱を選ぶ確率)
…となっている。もう一方も同様である。

で、問題は…
「置いてある箱からオレンジが取り出されたとき、それが赤い箱である確率」
…であった。

これはすなわち、以下の条件付き確率を求めたいということである。

$$p(B=r \mid F=o)$$

ベイズの定理より、先ほどの値を代入して計算すると…

$$p(B=r \mid B=o)=\frac{p(F=o \mid B=r)p(B=r)}{p(F=o)}=\frac{2}{3}$$

すなわち、「置いてある箱が赤い箱である確率」を考えたいとき、事前には単純な
「赤い箱が置いてある確率」(事前確率)
だけしか知らなかったが、その後果物(ここではオレンジであった)を取り出すことで
「オレンジが取り出されたときに、それが赤い箱である確率」(事後確率)
というふうに絞り込むことができるのである。

では、ベイズ確率とは？

◆古典的(頻度主義的)確率解釈、すなわち確率を「ランダムな繰り返し試行の頻度」とみなすだけでなく、より広義に「不確実性の度合い」とする解釈

◆たとえば「月がかつて太陽を周る軌道上にあったかどうか」「南極の万年雪が今世紀末には消えるかどうか」など、たくさんの繰り返しが観測できない事象は、頻度主義的な確率解釈では「確率」として捉えられない。

◆…しかしこれらの事象が「どのくらいの尤もらしさで起こる/起こったのであろうか」ということに関して我々は何らかの知見を持っているし、そこに新たなデータ(温室効果ガスについての観測衛星からの情報など)がつけ加われば、その尤もらしさについての知見を修正することもできる。

◆このように不確実性(や信念の度合い)を定量的に表現し、新な証拠に照らして修正し、またその結果として最適な行動や決定を下そうとする場合に有用なのがベイズ確率である。

…すなわち、データを観測する前の我々の仮説を事前確率分布

$$p(\mathbf{w})$$

として取り込んでおき、これを、観測することによって新たに得られたデータ

$$\mathcal{D} = \{t_1, \ldots, t_N\}$$

を用いて、新たなデータを照しあわせたときの事前分布の尤もらしさを
以下のような事後分布として評価し、修正できることになる。

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w}) p(\mathbf{w})}{p(\mathcal{D})}$$

ここで左辺の分子にある $p(\mathcal{D}|\mathbf{w})$ は、「**w というパラメータ**を仮定したときにデータ集合 $\mathcal{D}$ となるのはどのくらい尤もらしいことなのか」を表わす尤度関数である。

このようにベイズの定理は事後確率∝尤度×事前確率という形になっているが、頻度主義的アプローチとベイズ的アプローチではこの尤度関数の扱い方がおおきく異なっているといえる。

$$p(\mathbf{w}\,|\,\mathcal{D}) = \frac{p(\mathcal{D}\,|\,\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

◆頻度主義的アプローチでは、wは固定した、推定量として定められるパラメータとされ、データ集合 $\mathcal{D}$ はこのパラメータのもとでの分布に従った結果である。
◆ベイズ的アプローチでは実際に観測されたデータ集合 $\mathcal{D}$ がまずあって、そこから主観に基づくパラメータの不確実性が **w** の確率分布として表わされる。

たとえば「コインを投げて3回とも表が出た」場合…
◆頻度主義的アプローチでは「このコインは表が出る確率が1であるような確率分布に従ってるんだな」と考える。すなわち、データに対して尤度が最大となる **w** が固定される。
◆ベイズ的アプローチでは「はじめは表と裏が1/2ずつだと思っていたけれど、こんな結果が出たなら、その予想が当ってる度合いは低いだろう」と考える。すなわち、尤度によって **w** が評価され、$\mathcal{D}$ を条件とする分布に修正される。

…というように、ベイズ的アプローチは事前知識を自然に入れることができるため、頻度主義的アプローチのように極端な結論を導くことがない。

ベイズ確率の欠点

◆事前分布が何らかの信念によらず、むしろ数学的な便宜によって選ばれてしまうことがある。

◆事前分布の選び方によっては結果が主観的になるし、悪い事前分布を選べば、高い確率で悪い結果が得られてしまう。

◆頻度主義的アプローチを織り交ぜていくことで、ある程度回避することができる。

というわけで次回は今日やった曲線フィッティングをベイズ的にやります…