Choosing Predictors

CPS 271

Ron Parr

Regression figures provided by Christopher Bishop and © 2007 Christopher Bishop

What Makes a Good Prediction?

- Obviously: One that gives best performance in the future, but how do we pick this in advance?

- Best match to training set?
- Best match to training set (with regularization)?
- Distribution over hypotheses?
- Convergence to "truth" in the limit of infinite data?
- Data themselves + some interpolation rule?

Loss Functions

- Predict y, measure performance against target t
- One performance criterion is the squared loss:

$$E(y-t)^2$$

- Suppose we predict the mean, loss is then:

$$E(\bar{t}-t)^2$$

Expectation Minimize Loss

- Suppose you need to bet on an outcome (e.g. die roll)
- Suppose loss is squared error, want:

$$\min_y E(y-t)^2$$

- Minimize and solve for y

Sample Mean is Consistent

- Suppose we observe $X^{(1)} \ldots X^{(n)}$
- Assume these are independently drawn, and indentically distributed (IID)
- What is our estimate for E(X)?

$$\overline{X} = \frac{\sum_{i=1}^{n} X^{(i)}}{n}$$

- Why?

$$E(\overline{X}) = E\left(\frac{\sum_{i=1}^{n} X^{(i)}}{n}\right) = \frac{nE(X)}{n} = E(X)$$

Also…

Chebyshev's Inequality

- Let X have finite mean and variance:

$$P(|X - E(X)| \geq c) \leq \frac{Var(X)}{c^2}$$

- Variance governs our chances of missing the mean

Convergence of Sample Mean

- Apply Chebyshev's inequality to sample mean

$$P(|\overline{X} - E(\overline{X})| \geq c) \leq \frac{Var(\overline{X})}{c^2}$$

$$Var(\overline{X}) = Var\left(\sum_{i=1}^{n} \frac{X^{(i)}}{n}\right) = \sum_{i=1}^{n} \frac{1}{n^2} Var(X_i) = \frac{Var(X)}{n}$$

$$\lim_{n\rightarrow\infty} P(|\overline{X} - E(\overline{X})| \geq c) \leq \frac{Var(X)}{nc^2} = 0$$

Sample Variance

- Generalization of sample mean:

$$\overline{\sigma}^2 = \frac{\sum_{i=1}^{n} (x^{(i)} - \overline{x})^2}{n}$$

- Sample variance is biased:

$$E(\overline{\sigma}^2) = \sigma^2 \frac{n-1}{n}$$

Fitting Continuous Data (Regression)

- Datum i has feature vector: $\boldsymbol{\phi}=(\phi_1(x^{(i)})\ldots\phi_k(x^{(i)}))$
- Has real valued target: $t^{(i)}$
- Concept space: linear combinations of features:

$$y(\mathbf{x}^{(i)};\mathbf{w}) = \sum_{j=1}^{k} \phi_j(\mathbf{x}^{(i)}) w_j = \boldsymbol{\varphi}(\mathbf{x}^{(i)})^T \mathbf{w}$$

- Learning objective: Search to find "best" w
- (This is standard "data fitting" that most people learn in some form or another.)

Linearity of Regression

- Regression typically considered a *linear* method, but…
- Features not necessarily linear
- Features not necessarily linear
- Features not necessarily linear
- Features not necessarily linear
- and, BTW, features not necessarily linear

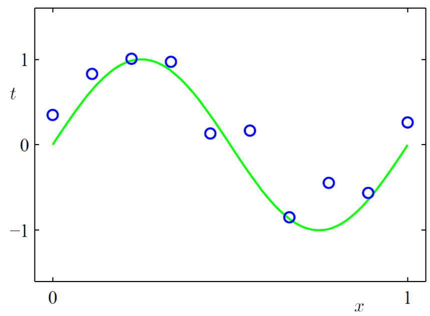
Regression Examples

- Predicting housing price from:
  - House size, lot size, rooms, neighborhood*, etc.
- Predicting weight from:
  - Sex, height, ethnicity, etc.
- Predicting life expectancy increase from:
  - Medication, disease state, etc.
- Predicting crop yield from:
  - Precipitation, fertilizer, temperature, etc.
- Fitting polynomials
  - Features are monomials

What Regression Does

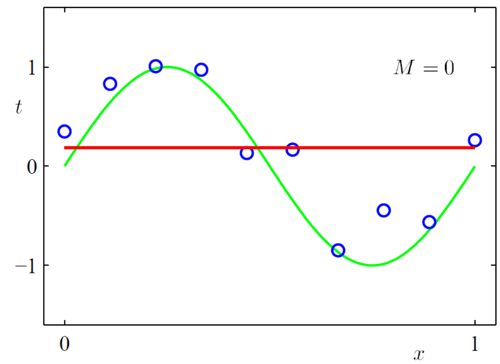
- Regression
  - Minimizes squared error on training set
  - Projects training set into linear subspace spanned by the features
- We will prove some of these properties later in the class

What is the Best Choice of Polynomial?

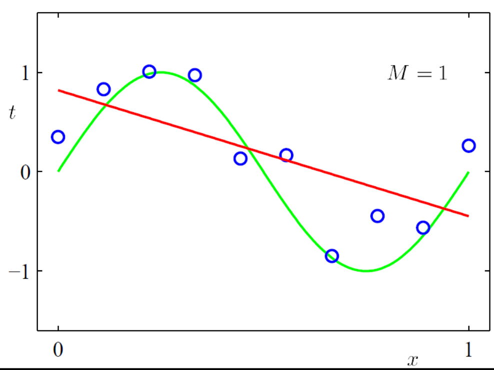


Noisy Source Data

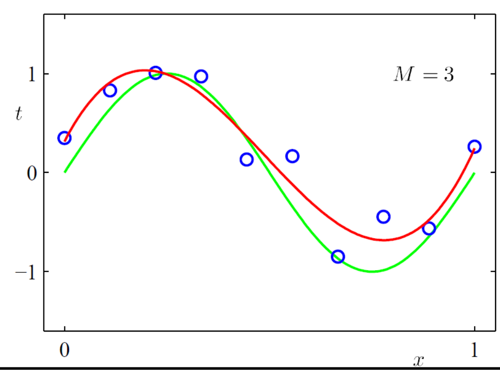
Degree 0 Fit



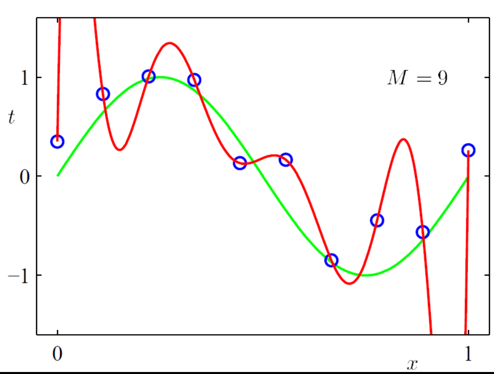
Degree 1 Fit



Degree 3 Fit

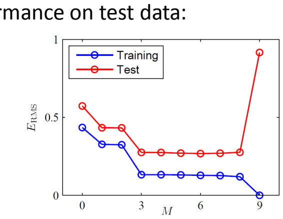


Degree 9 Fit



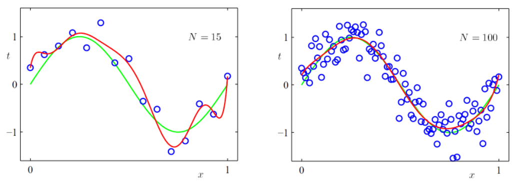
Observations

- Degree 3 is the best match to the source
- Degree 9 is the best match to the samples
- We call this **over-fitting**
- Performance on test data:



What went wrong?

- Is the problem a bad choice of polynomial?
- Is the problem that we don't have enough data?
- Answer: Yes



Regularization

- Idea: Penalize overly complicated answers
- Regular regression minimizes:

$$\sum_{i=1}^{M}(y(x^{(i)};\mathbf{w})-t_i)^2$$

- Regularized regression minimizes:

$$\lambda f(\|\mathbf{w}\|)+\sum_{i=1}^{M}(y(x^{(i)};\mathbf{w})-t_i)^2$$

- Note: May exclude constants form the norm

Regularization: Why?

$$\lambda f(\|\mathbf{w}\|)+\sum_{i=1}^{M}(y(x^{(i)};\mathbf{w})-t^{(i)})^2$$

- For polynomials, extreme curves typically require extreme values
- In general, encourages use of features only when they lead to a substantial increase in performance
- Problem: How to choose λ

A Bayesian Perspective

- Suppose we have a space of possible hypotheses H
- Which hypothesis has the highest posterior:

$$P(H \mid D)=\frac{P(D \mid H)P(H)}{P(D)}$$

- P(D) does not depend on H; maximize numerator
- Uniform P(H) is called Maximum Likelihood solution (model for which data has highest prob.)
- P(H) can be used for regularization

Maximum Likelihood

- For many models, the empirical mean is also the maximum likelihood solution
- Suppose:
  - Data normally distributed
  - Unknown mean, variance
  - IID samples

$$P(D \mid H)=P(t^{(1)}..t^{(m)} \mid \mu,\sigma)$$

$$=\prod_{i=1}^{m}\frac{e^{-\frac{(t^{(i)}-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

Maximum Likelihood for Gaussians

- Sample mean is ML solution for mean
- Sample variance is ML solution for variance

$$\mu_{ML}=\frac{\sum_{i=1}^{n}t^{(i)}}{n}$$

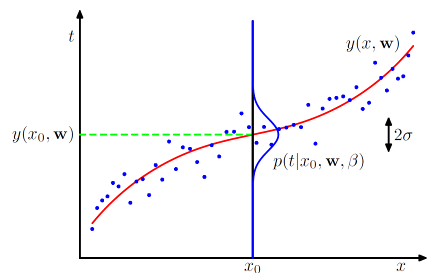
$$\sigma_{ML}{}^2=\frac{\sum_{i=1}^{n}(t^{(i)}-\mu_{ML})^2}{n}$$

Priors for Gaussians

- Recall Bayes rule: $P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}$
- Does it make sense to have a P(H) for Gaussians?
- Yes: Corresponds to some prior knowledge about the mean or variance
- Would like this knowledge to have a mathematically convenient form
- We will see later that the **Wishart** distribution is a *conjugate prior* for the Gaussian distribution w/known mean

Bayesian Regression

- Assume that, given x, noise is Gaussian
- Homoscedastic noise model



Maximum Likelihood Solution

$$P(D \mid H) = P(t^{(1)} \ldots t^{(m)} \mid y(x;\mathbf{w}), \sigma)$$

$$= \prod_{i=1}^{m} \frac{e^{\frac{-(t^{(i)} - y(x;w))^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

- ML fit for mean is just linear regression fit
- ML fit for mean does not depend upon σ

Bayesian Solution

- Introduce prior distribution over weights

$$p(H) = p(w \mid \alpha) = N(w \mid 0, \frac{1}{\alpha} I)$$

- Posterior now becomes:

$$P(D \mid H)P(H) = P(t^{(1)} \ldots t^{(m)} \mid y(x;\mathbf{w}), \sigma)P(\mathbf{w})$$

$$= \prod_{i=1}^{m} \frac{e^{\frac{-(t^{(i)} - y(x;w))^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \frac{e^{\frac{-\alpha \mathbf{w}^T \mathbf{w}}{2}}}{\frac{2\pi}{\alpha}^{(k+1)/2}}$$

Comparing Regularized Regression with Bayesian Regresion

- Regularized Regression minimizes:

$$\lambda f(\|\mathbf{w}\|) + \sum_{i=1}^{M} (y(x_i;\mathbf{w}) - t_i)^2$$

- Bayesian Regression maximizes:

$$\prod_{i=1}^{m} \frac{e^{\frac{-(t^{(i)} - y(x;w))^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \frac{e^{\frac{-\alpha \mathbf{w}^T \mathbf{w}}{2}}}{\frac{2\pi}{\alpha}^{(k+1)/2}}$$

- Observation: Take log of Bayesian regression criterion and these become identical (up to constants)

Regularization: An Empirical Approach

- Problem: We still have a magic constant that trades off complexity vs. fit
- Solution 1:
  - Generate multiple models
  - Use lots of test data to discover and discard bad models
- Solution 2 - *S-fold cross validation*:
  - Divide data into S groups
  - Create validation set i by subtracting group i form original set
  - Produces S groups of size (S-1)/S
  - Train on S-1, Test on held out set
  - Repeat, combine results in some way

Conclusions

- Many methods for choosing the best hypothesis – no single best w/o more information about the task
- Maximum likelihood and minimum squared error on training set are similar/same under some common assumptions
- Regularization prevents overfitting, is necessary when data are scarce