Least Squares Regression and Bias-Variance Decomp.

David Helmbold

University of California, Santa Cruz
dph@soe.ucsc.edu

Fall ’12, revised W’14, S’15, W’16

Regression

Given training data $\{(\mathbf{x}_i, y_i)\}_{i=1..m}$, find a linear approximation to the y's.

i.e. find $\mathbf{w}$ (or θ) of weights/parameters such that $\mathbf{w} \cdot \mathbf{x} \approx y$

Topics:

- Least Squares as Maximum likelihood
- Finding Maximum-likelihood weights
- Bias-variance error decomposition
- Basis functions for transforming features
- 1-norm and 2-norm Regularization
- Bayesian Linear Regression

Maximum Likelihood Regression

- Learn a function f in a given class of functions (not necc. linear)
- Have m examples $\{(\mathbf{x}_i, y_i)\}$ where $y_i = f(\mathbf{x}_i) + \underbrace{\epsilon_i}_{\text{noise}}$
- Assume $\mathbf{x}$'s fixed, concentrate on y's   like discriminative
- Assume ϵ_i's are iid draws from some mean 0 Gaussian distribution
- Probability of getting y_i for $\mathbf{x}_i$ with f is:

$$\begin{aligned} p(y_i \mid \mathbf{x}_i, f) &= p(\epsilon_i = y_i - f(\mathbf{x}_i)) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - f(\mathbf{x}_i))^2}{2\sigma^2}\right) \end{aligned}$$

- Likelihood of f: $\mathcal{L}(f) = P(\text{all labels} \mid f, \text{all } \mathbf{x}) = \prod_{i=1}^{m} p(y_i \mid \mathbf{x}_i, f)$
  Prob. of getting all the y_i's using f

$$
\begin{aligned}
\ln \mathcal{L}(f) &= \sum_{i=1}^{m} \ln p(y_i \mid \mathbf{x}_i, f) \\
\ln \mathcal{L}(f) &= m \ln \left(\frac{1}{\sqrt{2\pi}\sigma} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - f(\mathbf{x}_i))^2
\end{aligned}
$$

- To maximize likelihood of f, minimize the squared error!

back to linear regression ...

What $\mathbf{w}$ maximizes the likelihood?

- Consider

$$\nabla_{\mathbf{w}} \ln \mathcal{L}(\mathbf{w}) = \nabla_{\mathbf{w}} \left(m \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - \mathbf{w} \cdot \mathbf{x}_i)^2 \right)$$

$$= \left(\frac{1}{\sigma^2} \sum_{i=1}^{m} (y_i - \mathbf{w}^\top \mathbf{x}_i) \mathbf{x}_i^\top \right)$$

- set eq, to 0 and solve for $\mathbf{w}$

$$\mathbf{0}^\top = \sum_{i=1}^{m} y_i \mathbf{x}_i^\top - \mathbf{w}^\top \sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top$$

transpose and using matrix magic

$$\mathbf{w}_{ML} = (X^\top X)^{-1} X^\top \mathbf{y}$$

Where row i of X is instance $\mathbf{x}_i$.
$(X^\top X)^{-1} X^\top$ is the pseudo-inverse of X

Matrix Magic Example: $D =$ number of features

$$\sum_{i=1}^{m} y_i \mathbf{x}_i = \begin{pmatrix} \sum y_i \, x_{i,1} \\ \sum y_i \, x_{i,2} \\ \cdots \\ \sum y_i \, x_{i,D} \end{pmatrix}$$

$x_{i,j}$ is jth feature of example i

$$= \begin{pmatrix} x_{1,1} & x_{2,1} & \cdots & x_{m,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{m,2} \\ \vdots & \ldots & & \vdots \\ x_{1,D} & x_{2,D} & \cdots & x_{m,D} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

$$= \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,D} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,D} \\ \vdots & \ldots & & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,D} \end{pmatrix}^{\top} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = X^{\top} \mathbf{y}$$

Here $^{\top}$ means transpose

Can also use stochastic gradient descent to learn $\mathbf{w}$.

Cycle through examples, taking a step in the (negative) gradient direction for each example $(\mathbf{x}_n, y_n)$

$$
\begin{aligned}
\mathbf{w}_{\text{new}} &= \mathbf{w}_{\text{old}} - \eta \nabla \text{Error}(\mathbf{x}_n, y_n) \\
&= \mathbf{w}_{\text{old}} - \eta \nabla \frac{1}{2} (y_n - \mathbf{w}_{\text{old}} \cdot \mathbf{x}_n)^2 \\
&= \mathbf{w}_{\text{old}} + \eta (y_n - \mathbf{w}_{\text{old}} \cdot \mathbf{x}_n) \mathbf{x}_n
\end{aligned}
$$

Known as LMS algorithm
How to choose η?

Bias-Variance decomposition

Goal: enlightening breakdown of a regression function's expected error

Fix training instances $\mathbf{x}_1, \dots \mathbf{x}_m$ and "true" target function f

Let labels y_i in sample be $f(\mathbf{x}_i) + \epsilon_i$ where ϵ_i's are any iid noise.

Assume $E[\epsilon_i] = 0$, so $E[y_i] = f(\mathbf{x}_i)$.

We will examine the expected squared error between regression function g learned from sample and "true" function f at a particular test point $\mathbf{x}$.

$$E_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}))^2]$$

where E_{noise} is expectation over training label noise.

What is the experiment?

Let $\bar{g}(\mathbf{x})$ be the $E_{\text{noise}}(g(\mathbf{x}))$. We can re-write $E_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}))^2]$

$$
\begin{aligned}
&= E_{\text{noise}}\left[(g(\mathbf{x}) - \bar{g} + \bar{g} - f(\mathbf{x}))^2\right] \\
&= E_{\text{noise}}\left[(g(\mathbf{x}) - \bar{g})^2 + (\bar{g} - f(\mathbf{x}))^2 + 2(g(\mathbf{x}) - \bar{g})(\bar{g} - f(\mathbf{x}))\right] \\
&= E_{\text{noise}}\left[(g(\mathbf{x}) - \bar{g})^2\right] + E_{\text{noise}}\left[(\bar{g} - f(\mathbf{x}))^2\right] \\
&\qquad + 2E_{\text{noise}}\left[(g(\mathbf{x}) - \bar{g})(\bar{g} - f(\mathbf{x}))\right] \\
&= E_{\text{noise}}\left[(g(\mathbf{x}) - \bar{g})^2\right] + (\bar{g} - f(\mathbf{x}))^2 + 2E_{\text{noise}}\left[(g(\mathbf{x}) - \bar{g})(\bar{g} - f(\mathbf{x}))\right] \\
&= \text{variance of } g(\mathbf{x}) + (\text{bias of } g(\mathbf{x}))^2 + 2E_{\text{noise}}\left[(g(\mathbf{x}) - \bar{g})(\bar{g} - f(\mathbf{x}))\right]
\end{aligned}
$$

But $f(x)$ and $\bar{g}$ constant with respect to the noise, so

$$\begin{aligned}
E_{\text{noise}}\left[(g(\mathbf{x}) - \bar{g})(\bar{g} - f(\mathbf{x}))\right] &= (\bar{g} - f(\mathbf{x}))E_{\text{noise}}\left[g(\mathbf{x}) - \bar{g}\right] \\
&= (\bar{g} - f(\mathbf{x}))\left(E_{\text{noise}}\left[g(\mathbf{x})\right] - E_{\text{noise}}\left[\bar{g}\right]\right) \\
&= (\bar{g} - f(\mathbf{x}))\left(\bar{g} - \bar{g}\right) \\
&= 0
\end{aligned}$$

Taking expectation over **x**'s in training data similar.

Now look at squared error to $y = f(\mathbf{x}) + \epsilon$ on test point $\mathbf{x}$

$E_{\text{noise}}[(g(\mathbf{x}) - y)^2]$

$$\begin{aligned}
&= E_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}) + f(\mathbf{x}) - y)^2] \\
&\qquad \text{recall } y = f(\mathbf{x}) + \epsilon, \text{ so } f(\mathbf{x}) - y = -\epsilon \\
&= E_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}) + -\epsilon)^2] \\
&= E_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}))^2 + \epsilon^2 - 2\epsilon(g(\mathbf{x}) - f(\mathbf{x}))] \\
&= E_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}))^2] + E_{\text{noise}}[\epsilon^2] - 2E_{\text{noise}}[\epsilon(g(\mathbf{x}) - f(\mathbf{x}))] \\
&\qquad \epsilon \text{ and } g(\mathbf{x}) - f(\mathbf{x}) \text{ independent RVs so expectations multiply} \\
&= E_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}))^2] + E_{\text{noise}}[\epsilon^2] - 2E_{\text{noise}}[\epsilon]E_{\text{noise}}[g(\mathbf{x}) - f(\mathbf{x}))] \\
&= E_{\text{noise}}[(g(\mathbf{x}) - f(\mathbf{x}))^2] + E_{\text{noise}}[\epsilon^2] \\
&= \text{expected squared error to } f + \text{variance due to noise} \\
&= \text{variance of } g(\mathbf{x}) + (\text{bias of } g(\mathbf{x}))^2 + \text{variance due to noise}
\end{aligned}$$