03/02/2016

Statistical Justification for Regression

- Typical model of reality:
  - Samples come from unknown probability distribution $X_i \sim D$.
  - y-values are sum of unknowns, non-random surface plus random noise: for all X_i,

$$y_i = f(X_i) + \epsilon_i$$

- Goal of regression: find h that estimates f.
- Ideal approach: choose $h(x) = E_y[Y|X = x]$

Least-squares Regression from Max Likelihood

- Suppose $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$; then $y_i|X_i \sim \mathcal{N}(f(X_i), \sigma^2)$.
- Recall that log likelihood for normal distribution is,

$$lnP(y_i) = -\frac{|y_i - \mu|^2}{2\sigma^2} - constant \qquad \Leftarrow \mu = f(X_i)$$
$$ln(P(y_1)P(y_2)\dots P(y_n)) = lnP(y_1) + lnP(y_2) + \cdots + lnP(y_n)$$

- Takeaway: If you apply the principle of max likelihood to linear regression with an input model that assumes gaussian noise $\Rightarrow$ find f by least-squares.

Empirical Risk

- The risk for hypothesis h is the expected loss $R(h) = E[L]$ over all X, Y.
- Discriminative model: we don't know X's distribution D. How can we minimize the risk?
- Empirical distribution: A discrete probability that is the sample set, with each sample equally likely.
- Empirical risk: expected loss over empirical distribution $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(X_i), y_i))$.
- Takeaway: this is why we minimize the sum of loss functions.

Logistic regression from Max Likelihood

- If we accept the logistic regression function, what cost function should we use?
- Given arbitrary sample x, write probability it is in (not in) the class: (fictitious dimension: x ends w/1; w ends w/α).

$$P(y = 1|x; w) = h(x; w) \qquad \Leftarrow h(x; w) = s(w^T x)$$
$$P(y = 0|x; w) = 1 - h(x; w)$$

- Combine these 2 facts into 1 expression:

$$P(y|x; w) = h(x)^y (1 - h(x))^{1-y}$$

- Likelihood is,

$$\begin{aligned}
L(w; x_1, \ldots, x_n) &= \prod_{i=1}^{n} P(y_i|X_i; w) \\
l(w) = lnL(w) &= \sum_{i=1}^{n} lnP(y_i|X_i; w) \\
&= \sum (y_i lnh(X_i) + (1-y_i)ln(1-h(X_i)))
\end{aligned}$$

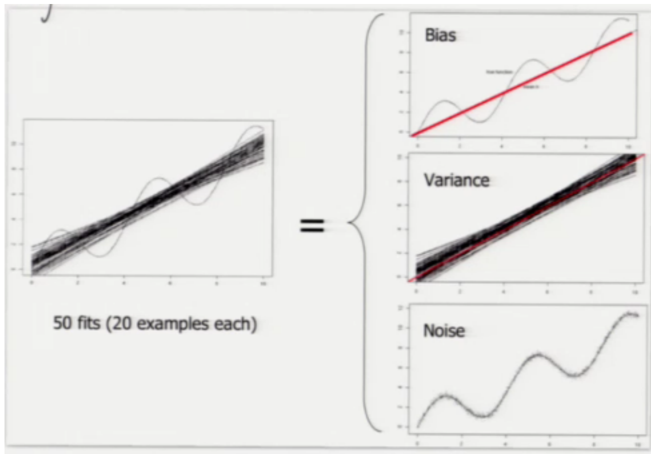
- which is negated logistic cost function J(w).

The Bias-variance Decomposition

- There are 2 sources of error in a hypothesis h:
  - bias: error due to inability of hypothesis h to fit f perfectly. e.g. fitting quadratic f with a linear ℓ.
  - variance: error due to fitting random noise in data. e.g. we fit linear f with a linear h, yes $h \neq f$.
- Model: generate samples $X_1 \ldots X_n$ from some distribution D. Values $y_i = f(X_i) + \epsilon_i$. Fit hypothesis h to X, y.
- Now h is a random variable; i.e. its weights are random.
- Consider an arbitrary point $z \in \mathbb{R}^d$ (not necessarily a sample!) and $\gamma = f(z) + \epsilon$.
- Note: $\mathrm{E}[\gamma] = f(z); \mathrm{Var}(\gamma) = \mathrm{Var}(\epsilon)$;
- Risk function when loss is squared error:
- Here we are taking the expectation over all possible training sets X, y and values of γ.

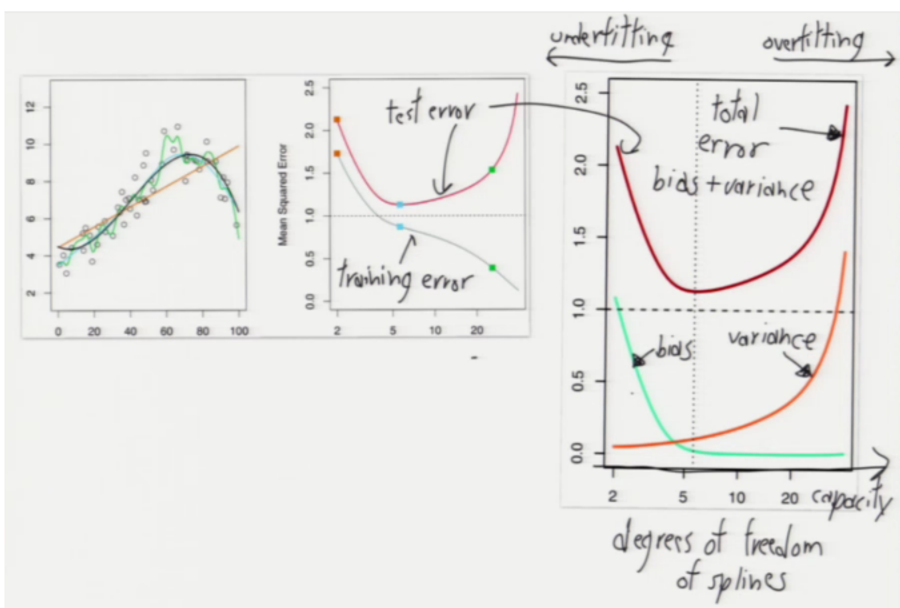
$$\begin{aligned}
R(h) &= E[L(h(z), \gamma)] \\
&= E[(h(z) - \gamma)^2] \\
&= E[h(z)^2] + E[\gamma^2] - 2E[\gamma h(z)] \\
&= \mathrm{Var}(h(z)) + E[h(z)]^2 + \mathrm{Var}(\gamma) + E[\gamma]^2 - 2E[\gamma]E[h(z)] \\
&= (E[h(z)] - E[\gamma])^2 + \mathrm{Var}(h(z)) + \mathrm{Var}(\gamma) \\
&= E[h(z) - f(z)]^2 + \mathrm{Var}(h(z)) + \mathrm{Var}(\epsilon)
\end{aligned}$$

  - We take expectation over possible training sets, X, y and values of γ.
  - $E[h(z) - f(z)]^2$: is square of the bias of method.
  - $\mathrm{Var}(h(z))$: variance of method.
  - $\mathrm{Var}(\epsilon)$: irreducible error (comes from test point not from training).
- This is point-wise version. Mean version: Let $z \sim D$ be random variable; take expectation of the squares bias, variance over z



  - Under-fitting: too much bias.
  - Over-fitting caused by too much variance.

- Training error reflects bias but not variance; test error reflects both.
- For many distributions, variance $\to 0$ as $n \to \infty$.
- If h can fit f exactly, for many distributions bias $\to 0$ as $n \to \infty$.
- If h cannot fit f well, bias is large at "most" points.
- Adding a good feature reduces bias; adding a bad feature rarely increases it.
- Adding a feature usually increases variance.
- Cannot reduce irreducible error.
- Noise in test set affects only var(ϵ); noise in training set affects only bias and Var(h).
- For real-world data, f is rarely knowable (and noise model might be wrong).
- But we can test learning algorithms by choosing f and making synthetic data.



- Example: Least-Squares Linear Regression:
  - No fictitious dimension.
  - Model: $f(z) = v^T z$.
  - Let e be a noise n-vector $e \sim \mathcal{N}(0, \sigma^2)$.
  - Training values: $y = Xv + e$. Input to regression algorithm are y, X.
  - Linear regression computes weights:

$$w = X^+ y = X^+(Xv + e) = v + X^+ e$$

  - Bias is,

$$E[h(z) - f(z)] = E[w^T z - v^T z] = E[z^T X^+ e] = z^T X^+ E[e] = 0$$

  - Warning: This does not mean $h(z) - f(z)$ is everywhere 0. Sometimes positive, sometimes negative, mean over training sets is 0.
  - Variance is,

$$\begin{aligned}
\mathrm{Var}(h(z)) &= \mathrm{Var}(w^T z) = \mathrm{Var}(z^T w) \\
&= \mathrm{Var}(z^T(v + X^+ e) \\
&= \mathrm{Var}(z^T v + z^T X^+ e)) \\
&= \mathrm{Var}(z^T X^+ e) \\
&= \sigma^2 |z^T X^+|^2 \\
&= \sigma^2 |z^T (X^T X)^{-1} X^T)|^2 \\
&= \sigma^2 z^T (X^T X)^{-1} X^T X (X^T X)^{-1} z \\
&= \sigma^2 z^T (X^+ X)^{-1} z
\end{aligned}$$

  - If choose coordinate system so $E[X] = 0$ it simplifies to $\approx \sigma^2 \frac{d}{n}$.
  - Takeaways: Bias can be zero when hypothesis function can fit the real one. Variance portion of RSS (overfitting) decreases as $\frac{1}{n}$, increases as d.