UNIVERSITY OF CALIFORNIA, BERKELEY

Lecture 12: CS 189 - Statistical Justifications for Regression

Oscar Ortega

July 16, 2021

1 TYPICAL MODEL OF REALITY:

- sample points come from unknown prob Distribution: $X_i = D$
- y-values are sum of unknown, non-random fn.s + random noise

$$\forall x_i, y_i = f(x_i) + \epsilon_i, e_i = D', \mathbb{E}(D') = 0$$

- goal of regression: find h that estimates f.

Ideal approach would be to choose $h(x) = \mathbb{E}_y(Y|X = x) = f(x) + \mathbb{E}(\epsilon) = f(x)$

2 LEAST- SQUARES REGRESSION FROM MAXIMUM LIKELIHOOD

Suppose $e_i = \mathcal{N}(0, \sigma^2)$, then $y_i = \mathcal{N}(f(X_i), \sigma^2)$
Recall that log likelihood for normal dist = $ln(P(y_i) = -\frac{y_i - \mu)^2}{2\sigma^2} - c, \leftarrow \mu = f(X_i)$

$$\ln(f; X, y) = \sum_i ln(P(y_i))$$

$$= -\frac{1}{2\sigma^2}\sum_i (y_i - f(x_i))^2 - c$$

Note, that the constant term is irrelevant to the maximization of this function.
Takeaway: max-likelihood on 'parameter' f that minimizes the sum of squared distances, or

in other words estimate f by performing least squares regression. However, take note than in instances where you can no longer assume Gaussian noise, changing up the regression model, with Maximum likelihood estimation, is a better bet.

2.1 Empirical Risk

The **risk** for hypothesis h is expected loss $R(H) = \mathbf{E}[L]$ over labels and outputs.
Discrimiative model: we don't know X's distribution, Hoe do we minimize risk?

Empirical Distribution: the **discrete** uniform distribution over the sample pts. With this definition, we can now define the **empirical risk** as the expected loss under the empirical distribution

$$\hat{R}(H) = \frac{1}{n}\sum_i L(h(X_i), y_i)$$

Takeaway: this is why we minimize the sum of loss functions.

2.2 Logistic Loss from Maximum Likelihood

What cost fn should we use should we use or probabilities? Actual probability pt X_i is in the class y_i; predicted probability is $h(x_i)$
Thought experiment:
imagine b duplicate copies of X_i, $y_i b$ are in the class, $(1-y_i)b$ are not.This would mean the likelihood is as follows:

$$\mathcal{L}(h; X, y) = Bin(h, y_i b + (1-y_i)b)$$

this implies the log-likelihood $= -b\sum_i(y_i ln(h(X_i + (1-y_i)ln(1-h(X_i)) = -b\sum_i$ logistic loss fn $L(h(X_i, y_i))$

3 The Bias-Variance Decomposition

There are 2 sources of error in a hypothesis:

- **bias**: error due to inability of hypothesis to fit f perfectly.
- **variance**: error due to fitting random noise in data: e.g we fit linear f with a linear h, yet h does not equal f.

Model: $X_i := D, e_i := D', y_i = f(X_i) + \epsilon_i$
fit hypothesis h to X,y now h is a random variable; i.e its weights are random.
Consider an arbitrafy pt $z \in \mathbb{R}^d$ and $\gamma = f(z) + \epsilon$ Note:

$$\mathbb{E}(\gamma) = f(Z); var(\gamma) = var(\epsilon)$$

Risk fn when loss = squared error:

$$R(H) = \mathbb{E}[L(h(z), \gamma)]$$

We can interpret this as taking the expectation over all possible training sets, X,y and values of γ

$$= \mathbf{E}(h(z)-\gamma)^2) = \mathbf{E}(h(z))^2 + \mathbf{E}(\gamma^2) - 2\mathbf{E}(\gamma h(z))$$

$$= var(h(z) + \mathbf{E}(h(z))^2 + var(\gamma) + \mathbf{E}(\gamma)^2 - 2\mathbf{E}(\gamma)\mathbf{E}(h(z))$$

$$= (\mathbf{E}(h(z)) - \mathbf{E}(\gamma)))^2 + var(h(z)) + var(\gamma)$$

Note that in our computation, we assumed the true distribution is independent of the noise ('used that when going from line 2 to line 3').
We define the different components of this expression as follows:

- $\mathbb{E}((h(z)-\gamma^2))$ is the **bias of method**
- var$(h(z))$ is the **variance of method**
- var(γ) is the **irreducible error**

- **under-fitting** corresponds to having high bias
- **variance** corresponds to having high variance
- training error reflects bias but not variance, test error reflects both
- for many distributions: variance goes to zero as the number of points approaches infinity.
- if h can fit f exactly, for many distributions bias approaches 0 as the number of points approaches infinity
- if h cannot fit f well, bias is large at most points
- adding a good feature reduces the bias, rarely increases it
- adding a feature usually increases the variance
- can't reduce irreducible error
- noise in test set affects only $var(\epsilon)$
- noise in training set affects only bias and $var(h)$
- for real-world data, f is rarely knowable
- but we can test learning algs by choosing f and making synthetic data