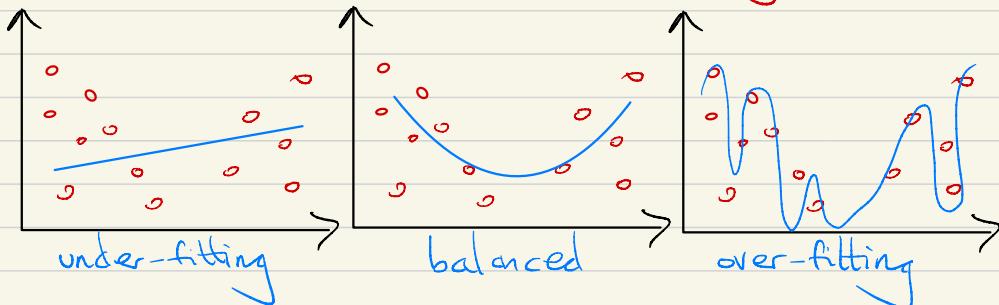


Lecture 18

03/24/2022

- Regularization

→ applying additional constraints to avoid over-fitting.



$$(Y - M\theta)^T C^{-1} (Y - M\theta) \rightarrow (Y - M\theta)^T C^{-1} (Y - M\theta) + (\lambda) \theta^T \theta$$

-- called RIDGE REGRESSION $\hat{\theta} = (M^T C^{-1} M + \lambda I)^{-1} M^T C^{-1} Y$

(NOTE) From a Bayesian perspective we have just applied a prior constraint on θ such that $p(\theta) \propto e^{-\frac{1}{2}\theta^T \theta}$

GENERALLY $\ln p(\theta) = -\frac{1}{2}(Y - M\theta)^T C^{-1} (Y - M\theta) - \frac{1}{2}\lambda \sum_{j=1}^n |\theta_j|^q$

$q=2$ (RIDGE REGRESSION) ... $q=1$ (LASSO REGRESSION)

6. Non-linear Regression

often we can make non-linear problems linear
e.g. $y = e^{\theta x} \Rightarrow \ln y = \theta x$

but not always ... Levenberg-Maquardt (LM) algorithm uses GRADIENT DESCENT and GAUSS-NEWTON OPTIMIZATION

e.g. expand $f(x_i | \theta)$ as TAYLOR EXPANSION --

$$f(x_i | \theta) = f(x_i | \theta_0) + \underbrace{J}_{\frac{\partial f(x_i | \theta)}{\partial \theta}} d\theta$$

LM minimizes sum of squares,
 $\sum [y_i - f(x_i | \theta_0) - J d\theta]^2$

... for a perturbation $d\theta$, that is updated by

$$[J^T C^{-1} J + \lambda \text{diag}(J^T C^{-1} J)] d\theta = J^T C^{-1} (Y - f(X | \theta))$$

λ small \rightarrow Gauss-Newton
 λ big \rightarrow gradient descent

damping parameter

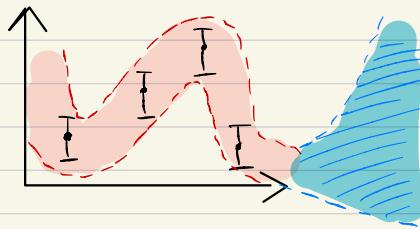
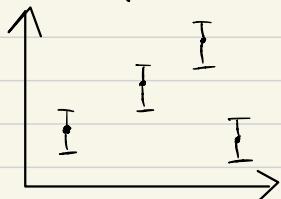
- Gaussian Process Regression (GPR)

↳ GPR treats your data vector as a big random sample taken from a N-D Gaussian distribution.

↳ by learning the covariance between points, GPR can interpolate --- (gives interpolation uncertainty too!)

[NOTE]

we are not drawing Gaussian shapes on the x-axis --- we model the data itself as a Gaussian draw.



↳ extrapolation can be poor.

- Total Least Squares ↳ uncertainties in "x" and "y"

