Lecture 8 Mar 11 2021

More on Generalization //

How well model does on new data

- no overfitting or underfitting

low bias
high variance

vs

high bias
low variance

How to control overfitting?

- Regularization (Weight Decay, L2)

Add factor of magnitude (norm) of coefficients of params to loss function so we minimize the magnitud of weights.

- Early Stopping

Take hyperparameters based on minimal validation error.

- Look at data distribution.

Cross-Validation //

- Split training data into 2 sets, validation and training

K-Fold Cross Validation

- split training data into k random subsets
- use k-1 subsets for training, remaining set for validation
- select model with smallest average validation error of k-folds

- This is costly since we must train and validate k times

- K-Fold allows us to have a bigger training set for better accuracy.

Leave-one-out Cross Validation

- K-Fold Cross Validation where k=N and N is size of training data
- somtimes we have small data sets so we must optimize for maximum training data

Problems with CV: Time consuming, not enough data, over/underfitting might still occure

Bayesian Methods //

Figuring out model parameters using probability.

Previous Methods:

- Maximum Likelihood $p(D|w)$

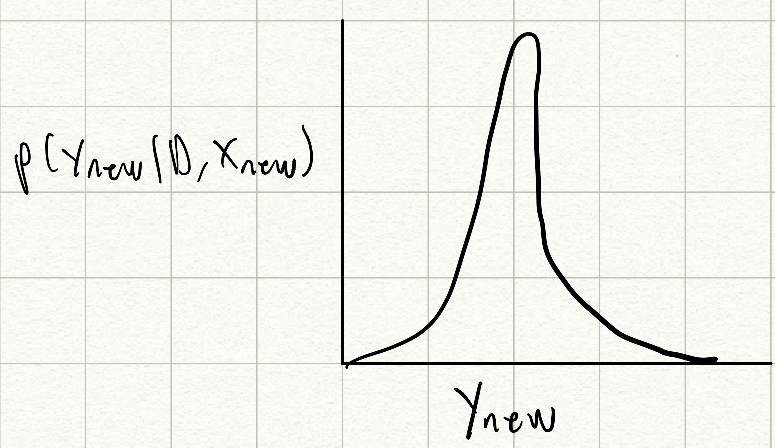
  This has significant chance that model will overfit.

- Maximum a Posteriori $p(w|D) = \frac{p(D|w)\,p(w)}{p(D)}$ ← constant

Less chance of overfitting since don't commit to a single model.

How to make predictions using $p(w|D)$ //

$p(Y_{new}|D, X_{new})$ gives us a PDF.



$$p(Y_{new}|D; x_{new}) = \int p(y_{new}|w, D, X_{new})\, p(w|D, X_{new})\, dw$$

probability of Y_{new} given model, data, X_{new}.
(Assume from Gaussian)

MAP distribution

$$= \int p(y_{new} \mid w, x_{new})\, p(w \mid D)\, dw$$

we usually use numerical methods to approximate this integral

This distribution tells us everything we need to know about beliefs on y_{new}.

☆ is the probability of y given data and x is equal to the weighted average of distributions $p(y \mid w, x)$, weighted by how well w fits the data $p(w \mid D)$ over all values of w.

Bayesian Regression //

With basis function, regression model is:

$$y = b(x)^T \vec{w} + \varepsilon \quad , \quad \varepsilon \sim N(0, \sigma^2)$$

$$b(x) = \begin{bmatrix} b_1(x) \\ b_2(x) \\ \vdots \\ b_m(x) \end{bmatrix}$$

The prior assumes $\vec{w} \sim (0, \alpha^{-1} I)$

The likelihood func is :

$$p(\vec{y} \mid \vec{x}, \vec{w}) = \prod^{N} p(y_i \mid x_i, \vec{w})$$

The posterior is :

$$p(\vec{w} \mid \vec{x}, \vec{y}) = \frac{\left(\prod^{N} p(y_i \mid x_i, \vec{w}) \right) p(w)}{\underbrace{p(\vec{y} \mid \vec{x})}_{p(D)}}$$

$$-\ln p(\vec{w} \mid \vec{x}, \vec{y}) = \frac{1}{2\sigma^2} \sum^{N} (y_i - f(x_i))^2 + \frac{\alpha}{2} \|w\|^2 + \underbrace{C}_{\text{constants}}$$

$$= \frac{1}{2\sigma^2} \|y - Bw\|^2 + \frac{\alpha}{2} \|w\|^2 + C$$

$$= \frac{1}{2} w^T \left(\frac{B^T B}{\sigma^2} + \alpha I \right) w - \frac{1}{2} \frac{y^T B w}{\sigma^2}$$

$$- \frac{1}{2} \frac{w^T B^T y}{\sigma^2}$$

$$+ C$$

$$= \frac{1}{2}(w-\bar{w})^T K^{-1}(w-\bar{w}) + c$$

$$\bar{w} = \frac{KB^T y}{\sigma^2} \qquad K = \left(\frac{B^T B}{\sigma^2} + \alpha I\right)^{-1}$$

This tells us the posterior distribution

$$p(\vec{w} \mid \vec{x}, \vec{y}) \sim N(w; \bar{w}, K)$$

Bayesian Prediction //

For new $x = x_{new}$

$$p(y_{new} \mid x_{new}, D) = \int p(y_{new} \mid w, D, x_{new})\, p(w \mid D)\, dw$$

$$= N(y_{new}; b(x_{new})^T w, \sigma^2 + b(x_{new})^T K b(x_{new}))$$

This predictive dist is a func from x_{new} to y_{new}.