ANLY-601

Advanced Pattern Recognition

Assignment Project Exam Help

Spring 2018

https://powcoder.com

Add WeChat powcoder

L11 – Map Estimates, Bayesian Inference, Hyperparameter Choice

GEORGETOWN UNIVERSITY

Continuing with Bayesian Methods

Assignment Project Exam Help

MAP Estimates, Bayesian Inference, and Hyperparameter Choice

https://powcoder.com

Add WeChat powcoder

Why use a MAP Estimate, they're Biased?

Consider the expected squared error of any estimator:

$$MSE = E\left[(\bar{\mu}-\mu)^2\right] = E\left[\left\{(\bar{\mu}-E[\bar{\mu}]) + (E[\bar{\mu}]-\mu)\right\}^2\right]$$

Assignment Project Exam Help

https://powcoder.com

$$= \quad E\left[(\bar{\mu}-E[\bar{\mu}])^2\right] + (E[\bar{\mu}]-\mu)^2$$

Add WeChat powcoder

$$= \quad \mathrm{var}(\bar{\mu}) + bias^2$$

Bias isn't the only consideration - variance is also important. There's usually a trade-off; increase the bias, and the variance drops, and vice-versa.

Bias-Variance Trade-Off and MAP Estimates

Let's go back to our MAP estimate of the mean for Gaussian data:

$$\bar{\mu} = \frac{m\lambda^2}{m\lambda^2 + \sigma^2} \frac{1}{m} \sum_{i=1}^{m} x_i + \frac{\sigma^2}{m\lambda^2 + \sigma^2} \mu_0$$

Assignment Project Exam Help
https://powcoder.com

The bias and variance are (show these!)

Add WeChat powcoder

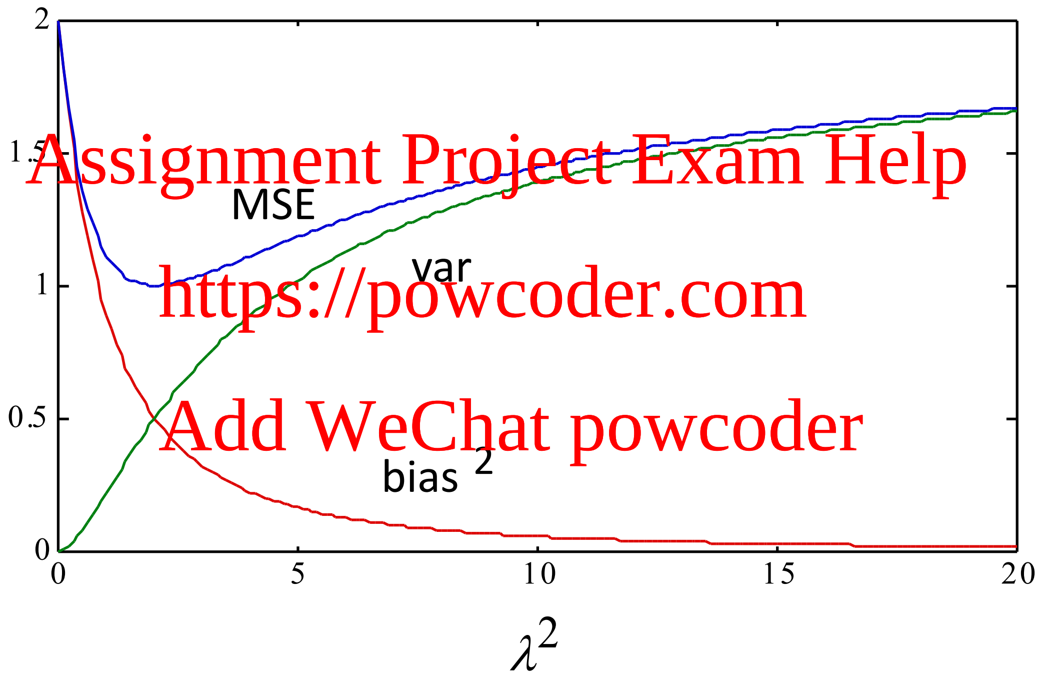
$$bias^2 = (E[\bar{\mu}] - \mu)^2 = \left(\frac{\sigma^2}{m\lambda^2 + \sigma^2} (\mu_0 - \mu) \right)^2$$

$$\mathrm{var}(\bar{\mu}) = E\left[(\bar{\mu} - E[\bar{\mu}])^2 \right] = \left(\frac{m\lambda^2}{m\lambda^2 + \sigma^2} \right)^2 \frac{\sigma^2}{m}$$

As $m \rightarrow \infty$ both go to zero

Bias-Variance Trade-Off and MAP Estimates

The curves look like this



The curve of MSE has its minimum at
a non-zero value of λ. Specifically -- $\lambda_{opt}{}^2 = (\mu_0 - \mu)^2$

GEORGETOWN UNIVERSITY

MAP Estimates and Regularizers

The log of the posterior on the parameters is

$$\log\left(p(\Theta \mid D)\right) = \log\left(p(D \mid \Theta)\right) + \log\left(p(\Theta)\right) - \log(p(D))$$

Assignment Project Exam Help

log-likelihood – bare cost

https://powcoder.com

log prior -- regularizer

Add WeChat powcoder

We saw that maximizing the data log-likelihood is equivalent to minimizing some nice cost function -- e.g. the mean-squared-error.

Maximizing the log-posterior is equivalent to minimizing a regularized cost function. **The effect of the regularizer is to reduce the parameter variance at the cost of adding parameter bias**.

MAP Regression

One can use the MAP estimate of Θ, and construct the regression function

$$E[t \mid x, D] = f(t \mid x, \hat{\Theta})$$

Assignment Project Exam Help

where $\hat{\Theta}$ is the value that maximizes the posterior p(Θ|D).

https://powcoder.com

One can also use this MAP value to estimate the target density

Add WeChat powcoder

$$p(t \mid x, \hat{\Theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{1}{2\sigma^2}\left(t - f(t \mid x, \hat{\theta})\right)^2$$

Example of Map Regression – Ridge

Ridge regression uses a parameterized regressor $f(x,\Theta)$, the familiar SSE cost function (Gaussian likelihood for the targets), and a Gaussian prior on the parameters, typically centered at zero

$$p(\theta) = \frac{1}{\sqrt{2\pi / \Lambda^2}} \exp -\frac{1}{2/\Lambda^2} |\theta|^2$$

Assignment Project Exam Help

https://powcoder.com

The regularized cost function is thus

Add WeChat powcoder

$$E(\Lambda, \theta) = \sum_{i=1}^{N} \left(t_i - f(x_i, \theta)\right)^2 + \Lambda |\theta|^2$$

That for linear regression, $f(x,\Theta)$ is linear in Θ so E is quadratic in Θ and the cost function can be minimized in closed form (just like MLE estimation for linear regression).

Bayesian Estimation

Let's continue. Suppose we have obtained the posterior on the parameters p(Θ| D) and we wish to find the probability of a new data value x. A Bayesian says that you should calculate this from his version of the distribution p(x)

Assignment Project Exam Help

$$p(x|D) = \int p(x\,|\,\Theta)\; p(\Theta\,|D)\, d\Theta$$

https://powcoder.com

A Bayesian computes the mean of any function f(x) as

Add WeChat powcoder

$$E\left[f\mid D\right] = \int f(x)\; p(x\,|\,D)\; dx = \iint f(x)\; p(x\,|\,\Theta)\; p(\Theta\,|\,D)\; d\Theta\; dx$$

Bayesian and MAP Estimates

Relation to **MAP** Estimates: Suppose the posterior is sharply peaked up about its maximum value (the MAP estimate). Write a series expansion of p(x|Θ) about the maximum and substitute into the integral

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

$$p(x\,|\,D) = \int p(x\,|\,\Theta)\, p(\Theta\,|\,D)\, d\Theta = \int \left[p(x\,|\,\hat{\Theta}) + \left.\frac{dp(x\,|\,\Theta)}{d\Theta}\right|_{\hat{\Theta}} (\Theta - \hat{\Theta}) \right.$$

$$\left. + \frac{1}{2} \left.\frac{d^2 p(x\,|\,\Theta)}{d\Theta^2}\right|_{\hat{\Theta}} (\Theta - \hat{\Theta})^2 + \ldots \right] p(\Theta\,|\,D)\, d\Theta$$

$$= p(x\,|\,\hat{\Theta}) + \left.\frac{dp(x\,|\,\Theta)}{d\Theta}\right|_{\hat{\Theta}} E\left[(\Theta - \hat{\Theta})\,|\,D\right] + \frac{1}{2} \left.\frac{d^2 p(x\,|\,\Theta)}{d\Theta^2}\right|_{\hat{\Theta}} E\left[\left(\Theta - \hat{\Theta}\right)^2 |\,D\right] + \ldots$$

Bayesian and MAP Estimates

$$p(x|D) = \int p(x|\Theta)\, p(\Theta|D)\, d\Theta = \int \left[p(x|\hat{\Theta}) + \left.\frac{dp(x|\Theta)}{d\Theta}\right|_{\hat{\Theta}} (\Theta - \hat{\Theta}) + \frac{1}{2} \left.\frac{d^2 p(x|\Theta)}{d\Theta^2}\right|_{\hat{\Theta}} (\Theta - \hat{\Theta})^2 + \ldots \right] p(\Theta|D)\, d\Theta$$

Assignment Project Exam Help

https://powcoder.com

$$= p(x|\hat{\Theta}) + \left.\frac{dp(x|\Theta)}{d\Theta}\right|_{\hat{\Theta}} E\left[(\Theta - \hat{\Theta})|D\right] + \frac{1}{2} \left.\frac{d^2 p(x|\Theta)}{d\Theta^2}\right|_{\hat{\Theta}} E\left[\left(\Theta - \hat{\Theta}\right)^2 |D\right] + \ldots$$

Add WeChat powcoder

Handwaving arg: As data increases, the posterior becomes more sharply peaked about the MAP value $\hat{\Theta}$, trailing terms will become small & integral is approximately

$$p(x|D) \approx p(x|\hat{\Theta})$$

Recursive Bayesian Estimation

Back to Bayesian estimation of p(x|D)

$$p(x|D) = \int p(x|\Theta)\, p(\Theta|D)\, d\Theta = \int p(x|\Theta) \frac{p(D|\Theta)\, p(\Theta)}{\int p(D|\Theta')\, p(\Theta')\, d\Theta'}\, d\Theta$$

Assignment Project Exam Help

https://powcoder.com

Denote the dataset with n points by D^n = {x1, x2, …, xn}, and its likelihood by

Add WeChat powcoder

$$p(D^n|\Theta) = \prod_{k=1}^{n} p(x_k|\Theta) = p(x_n|\Theta)\; p(D^{n-1}|\Theta)$$

Using the last expression, the posterior can be written

$$p(\Theta|D^n) = \frac{p(x_n|\Theta)\; p(\Theta|D^{n-1})}{\int p(x_n|\Theta')\; p(\Theta'|D^{n-1})\, d\Theta'}$$

Recursive Bayesian Estimation

We have written the posterior density for the n-sample data set as

$$p(\Theta \mid D^n) = \frac{p(x_n \mid \Theta)\ p(\Theta \mid D^{n-1})}{\int p(x_n \mid \Theta')\ p(\Theta' \mid D^{n-1}) d\Theta'}$$

Assignment Project Exam Help

Starting with zero data, we take $p(\Theta \mid D^0) = p(\Theta)$ and generate the sequence

https://powcoder.com

Add WeChat powcoder

$$p(\Theta \mid D^1),\ p(\Theta \mid D^2),\ \ldots$$

and thus <u>incrementally refine</u> our estimate of the posterior density as more and more data becomes available.

GEORGETOWN UNIVERSITY

How Does a Bayesian do Regression?

Get a dataset $D \equiv \{ (x_i, t_i) \; i = 1, \ldots, m \}$

Choose a parameterized regression function f(x;Θ) to fit to the data.

Choose a model distribution function for the targets, e.g. a Gaussian

Assignment Project Exam Help

https://powcoder.com

$$p(t \mid x, \sigma^2, \Theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} \left(t - f(x;\Theta) \right)^2 \right]$$

Add WeChat powcoder

Choose a prior distribution on the parameters $p(\Theta, \sigma^2)$.

Calculate the data likelihood and the posterior distribution of the parameters

$$p(\Theta, \sigma^2 \mid D) = \frac{1}{p(D)} \; p(D \mid \Theta, \sigma^2) \; p(\Theta, \sigma^2)$$

How Does a Bayesian Do Regression?

Calculate the target density as a function of x by integrating over the posterior distribution of the parameters

$$p(t \mid x, D) \;=\; \int p(t \mid x, \sigma^2, \Theta)\;\; p(\sigma^2, \Theta \mid D)\;\; d\sigma^2 \;\; d\Theta$$

From the distribution on t, we can calculate several quantities.

Assignment Project Exam Help

- The conditional mean E[t | x,D]) (called the <u>regressor</u>, and equal to

https://powcoder.com

$$E[t \mid x, D] = \int t \;\; p(t \mid x, D)\, dt$$

Add WeChat powcoder

$$= \int t\; p(t \mid x, \sigma^2, \Theta)\; dt\; d\sigma^2\; d\Theta = \int f(x; \Theta)\; p(\Theta \mid D)\, d\Theta$$

  for our Gaussian model.)

- The most likely value(s) of t $\quad \arg\max_t \; p(t \mid x, D)$

- The target variance var(t | x,D).

Hyperparameters and Model Selection

Our prior on model parameters is itself a parameterized distribution. Recall for our Gaussian density model, we put a prior on the distribution of the mean

Assignment Project Exam Help

https://powcoder.com

$$p(\mu \mid \mu_0, \lambda^2) = \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left[-\frac{1}{2\lambda^2} (\mu - \mu_0)^2 \right]$$

Add WeChat powcoder

But how were the <u>hyperparameters chosen</u> $\mu_0, \ \lambda^2$

Hyperparameter Selection

- We could calculate the likelihood function for particular values

$$p(D \mid \mu_0, \lambda^2) = \int p(D \mid \mu) \; p(\mu \mid \mu_0, \lambda^2) \; d\mu$$

and choose the values of the hyperparameters that maximizes it.

Assignment Project Exam Help

- We could set up a hyperprior on the hyperparameters and choose maximum aposteriori values for the hyperparameters by maximizing

https://powcoder.com

$$p(\mu_0, \lambda^2 \mid D) \propto p(D \mid \mu_0, \lambda^2) \; p(\mu_0, \lambda^2)$$

Add WeChat powcoder

(but the hyperprior is going to have its own parameters …).

- Use some sort of empirical technique.

Empirical Hyperparameter Selection

Using a ‘validation’ set and MAP estimates.
Divide data into two pieces, development and evaluation

Development	Eval.

Further divide the development set into fitting and validation

Assignment Project Exam Help

$\Lambda 1$	Fitting	Valid.
$\Lambda 2$	Fitting	Valid.
$\Lambda 3$	Fitting	Valid.

https://powcoder.com

Sweep the hyperparameters over a range $a<\Lambda<b$, and for value, fit the model finding the MAP estimate of the model parameter $\Theta(\Lambda_i,D)$. Measure the performance on the validation set.

Add WeChat powcoder

Find the Λ that gives the best performance on the validation set, and use this value to generate a new MAP estimate using the <u>entire</u> dev. set. Test on eval.

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

GEORGETOWN UNIVERSITY