jupyter review Last Checkpoint: 1 minute ago

File Edit View Run Kernel Nbgrader Settings Help

Markdown Ju

Final exam review

What did we learn in this course?

1. Python basics:
   - Python_1.ipynb
   - Python_2.ipynb
2. Numpy basics:
   - NumpyTutorial.ipynb
3. Linear Regression by vector derivatives:
   - LinearModels.ipynb,
   - PlaneFitProblem,
   - Hessians,
   - Practice Problems for Midterm 1,
4. Optimization by Gradient descent:
   - ContinuousOptimization.ipynb
5. 1-Layer Neural Network: Perceptron3.ipynb
6. Pre-midterm review: Practice Problems

minimize
Quadratic form
of vectors
Any vector expression
Quadratic
take its derivative
and equate it to zero

$$\underbrace{x^T Q x} + b^T x + c$$

Decision Theory:

$$\frac{\partial}{\partial x}\Big[\to x^T(Q+Q^T) + b^T + 0 = 0$$

$$\to x^T = -b^T(Q+Q^T)^{-1}$$

$$x = -(Q+Q^T)^{-T} b$$

- Lecture notes: decision-theory.pdf. Additional resources:
- Chapter 2 of MLStory book

Bayes Rule

Question:

Given two random variables X and Y specify the relationship between $P(X|Y)$ and $P(Y$

Answer

$P(X|Y)P(Y)$

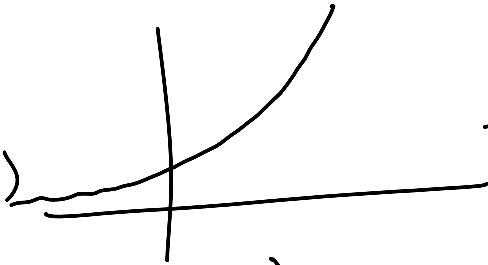
Regularization can be interpretted as application for Bayes theorem (Maximum-a-posterior estimate, MAP estimate)

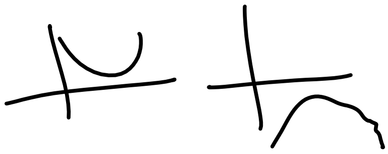
$$\arg\min_W L(D;W) + \lambda\|W\|_2^2 \iff \arg\max_W \underbrace{P(W|D)}_{\in[0,1]}$$

$$\arg\max_W \underbrace{-L(D;W) - \lambda\|W\|_2^2}_{\in(-\infty,\infty)}$$

In optimization, the optimal value/argument stays unchanged if you apply an increasing function. If you apply a decreasing function, the max changes to min and min changes to max.

$\exp(x) \in (0,\infty)$ when $x \in (-\infty,\infty)$





$$\arg\max_W \underbrace{\frac{1}{Z}\exp\left(-L(D;W) - \lambda\|W\|_2^2\right)}_{P(W|D)\ \text{posterior}} \quad \in[0,\infty)$$

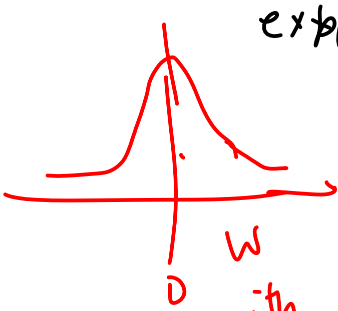
where Z is a normalization factor so that

$$\sum_W \exp(\ldots) = Z$$

$\exp(A+B)$

$\exp(A)\exp(B)$

$$\arg\max_W \frac{1}{\underbrace{Z}_{\text{Evidence}}} \underbrace{\exp\left(-L(D;W)\right)}_{\underbrace{P(D|W)}_{\text{likelihood}}} \overbrace{\underbrace{\exp\left(-\frac{\|W\|^2}{1/\lambda}\right)}_{\underbrace{P(W)}_{\text{Prior}}}}^{\text{Gaussian dist}}$$



with STD $\cdot\ \frac{1}{\lambda} = 2\sigma^2$

$\Rightarrow \sigma = \frac{1}{2\sqrt{\lambda}}$

Vector Jacobian product

Reverse mode differentiation

And it assumes the final output of the computation graph is a $\underline{\text{scalar}}$, (true for a loss function)

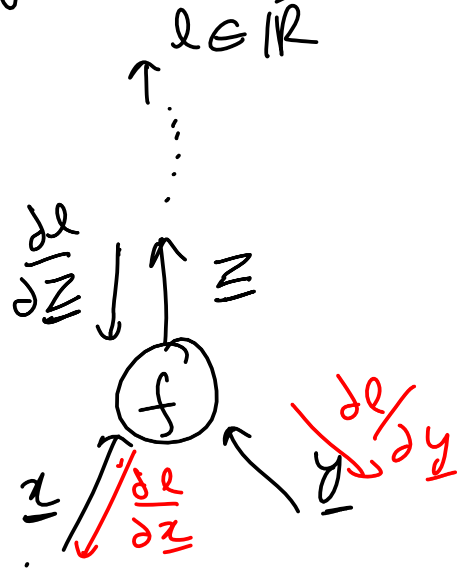
$$\frac{\partial l}{\partial x} = \frac{\partial l}{\partial z}\frac{\partial f}{\partial x} \leftarrow \text{Jacobian}$$

The responsibility of VJP function is to compute

$\frac{\partial l}{\partial x}$, $\frac{\partial l}{\partial y}$ given $\frac{\partial l}{\partial z}$ ← vector

and you can use x, y

$f(x, y)$



Examples

$$\underbrace{f(x, y) = I_{n\times n}\, x + y}$$

Find the VJP

Assume an eventual scalar function

$$\frac{\partial}{\partial x} l(f(x, y)) = \frac{\partial l}{\partial f}\frac{\partial f}{\partial x} = \frac{\partial l}{\partial f}\left(I_{n\times n} + 0_{n\times n}\right)$$

$$= \frac{\partial l}{\partial f} = \frac{\partial l}{\partial z}$$

$$\frac{\partial}{\partial y} l(f(x+y)) = \frac{\partial l}{\partial f} = \frac{\partial l}{\partial z}$$

$$\frac{\partial}{\partial x} Ax = A'$$

$$\frac{\partial}{\partial x}\begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix} x = \begin{bmatrix} a_1^T \\ \vdots \\ a_n^T \end{bmatrix}$$

$f(\underline{x}, \underline{y}) = \underline{x}^T \underline{y} \in \mathbb{R} \quad l \in \mathbb{R}$

$$\frac{\partial}{\partial \underline{x}} l(f(\underline{x}, \underline{y})) = \frac{\partial l}{\partial f} \frac{\partial f}{\partial \underline{x}} = \underbrace{\frac{\partial l}{\partial f} \underline{y}^T}$$

$$\frac{\partial}{\partial \underline{y}} l(f(\underline{x}, \underline{y})) = \frac{\partial l}{\partial f} \cdot \frac{\partial f}{\partial \underline{y}} = \frac{\partial l}{\partial f} \underline{x}^T$$

$f(x) = \sin(x)$

VJP?

$f(x) = \dfrac{1}{1 + \exp(-x)}$

VJP?

Project Convolution ×

$f(W, I) = W \otimes I$

VJP?

$f(\underline{x}, \underline{w}) = \dfrac{1}{1 + \exp(-\underline{w}^T \underline{x})}$

VJP?