1 Math

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

Set constraint $g(\mathbf{x})$ to zero and add multiplier.
$P(A_n, \ldots, A_1) = \mathrm{P}(A_n | A_{n-1}, \ldots, A_1) \cdot \mathrm{P}(A_{n-1}, \ldots, A_1)$
Joint: $\mathrm{P}\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} \mathrm{P}\left(A_k \,\middle|\, \bigcap_{j=1}^{k-1} A_j\right)$
Posterior: unobserved θ, observed x: $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$
Prior: prior belief in dist. of θ: $p(\theta)$
Likelihood: prob of observed given parameters: $p(x|\theta)$
MAP: $h_{MAP} = \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)P(h)$

2 todo

- Marginals
- Eigendecomposition
- Matrix multiplication
- Important derivatives
- Chain rule
- Product rule
- Log properties
- Positive semidefinite
- Convexity
- Def'n of norm from Bishop
- Polynomial multiplication
- Matrix transpose and inverse tproperties
- Gradient vs partial derivative

3 Regression

3.1 L2 regularization

Weights do not reach zero. Faster.

$$J_w = \frac{1}{2}(\Phi\mathbf{w} - \mathbf{y})^T(\Phi\mathbf{w} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

$$\mathbf{w} = (\Phi^T\Phi + \lambda\mathbf{I})^{-1}\Phi^T y$$

3.2 L1 regularization

Some weights set to zero. More expensive.

3.3 Gradient Descent

$\mathbf{w} \leftarrow \mathbf{w} - \alpha\nabla\log L(\mathbf{w})$

3.4 Bayesian regularization

4 Kernels

- Kernel properties

- Proving kernels

5 SVM

Minimize absolute error. More robust to outliers.