Review sheet

Derivatives: Rules on my slides
Logs: Rules on my slides
Linear algebra: $\mathbf{a}^T\mathbf{b}$ – also known as dot product of **a** and **b** or projection of **b** onto **a**, magnitude of **a**

Probability: Conditional, joint, marginal, Bayes rule $P(A|B) = \frac{P(B|A)P(A)}{P(B)} \propto P(B|A)P(A)$

Gaussian distribution

Training and testing sets, training and testing error

Performing classification: take class that has highest probability (Bayes) or "probability" (logistic regression), or based on whether binary classifier output is positive or negative (SVM)

Generative classifiers:
Binary coin flips:
Likelihood: $p^{|H|}(1-p)^{|T|}$
Prior: Beta distribution – role of α and β parameters
MLE: $p = \frac{|H|}{|H|+|T|}$
MAP: $p = \frac{|H|+\alpha-1}{|H|+|T|+\alpha-1+\beta-1}$

Multi-variate binary classification:
θ_{ik}^{j} is probability i^{th} feature has value x_k, given the class is y_j: $P(X_i=x_j|Y=y_j)$
θ^{j} is probability class is y_j : $P(Y=y_j)$

MLE: $P\left(X_i = x_k \middle| Y = y_j; \theta_{ik}^{j}\right) = \frac{\#D(X_i=x_k \wedge Y=y_j)}{\#D(Y=y_j)}$

$P\left(Y = y_j; \theta^{j}\right) = \frac{\#D(Y=y_j)}{|D|}$

MAP: $P\left(\theta_{ik}^{j} | X_i = x_k, Y = y_j\right) = \frac{\#D(X_i=x_k \wedge Y=y_j)+(\beta_j-1)}{\#D(Y=y_j)+\sum_m(\beta_m-1)}$

$P\left(\theta^{j} \; | Y = y_j\right) = \frac{\#D(Y=y_j)+(\beta_j-1)}{|D|+\sum_m(\beta_m-1)}$

Counting parameters:
Bayes with binary variables and two classes: $2\mathrm{x}(2^n-1)$
Naïve Bayes with binary variables and two classes: 2xn

Sigmoid: $g(h) = \frac{1}{1+e^{-h}}$

Logistic regression: Classes 0 and 1 (or multi-class, see below)

$$p(y=1|x;\theta) = g(\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1+e^{-w^T x}}$$
$$p(y=0|x;\theta) = 1 - g(\boldsymbol{w}^T\boldsymbol{x}) = \frac{e^{-w^T x}}{1+e^{-w^T x}}$$

MLE update:
For each data point (y^i,$\mathbf{x}^i$), update each weight w_j:
$$w_j \leftarrow w_j + \varepsilon x_j^i\left(y^i - g(\boldsymbol{w}^T\boldsymbol{x}^i)\right)$$

MAP update:
L1: $w_j \leftarrow w_j + \varepsilon\left[x_j^i\left(y^i - g(\boldsymbol{w}^T\boldsymbol{x}^i)\right) - \frac{\mathrm{sign}(w_j)}{\lambda N}\right]$
L2: $w_j \leftarrow w_j + \varepsilon\left[x_j^i\left(y^i - g(\boldsymbol{w}^T\boldsymbol{x}^i)\right) - \frac{w_j}{\lambda N}\right]$

Multi-class logistic regression:
For each data point (y^i,x^i), update each weight w_j:
$$w_{j,k} \leftarrow w_{j,k} + \varepsilon x_j^i\left(\delta(y^i - k) - p\left(y^i = k \middle| \boldsymbol{x}^i;\theta\right)\right)$$
$$\delta(a) = \begin{cases} 1 & \text{if } a = 0 \\ 0 & \text{otherwise} \end{cases}$$
Number of parameters: M classes, N features – (N+1)x(M-1) regression parameters

Support vector machines: Classes -1 and +1

Maximize margin, minimize $\mathbf{w}^T\mathbf{w}$ subject to $\boldsymbol{w}^T\boldsymbol{x}^+ + b \geq 1$ and $\boldsymbol{w}^T\boldsymbol{x}^- + b \leq -1$

Solution to max-margin problem: $w = \sum_i \alpha^i\ \boldsymbol{x}^i y^i \quad \alpha^i \geq 0\ \forall i \quad \sum_i \alpha^i\ y^i = 0$

Slack variables – penalizing errors

$\text{argmin}_{w,b}\ \boldsymbol{w}^T\boldsymbol{w} + C\sum_i \varepsilon_i$ subject to $\boldsymbol{w}^T\boldsymbol{x}^+ + b \geq 1 - \varepsilon_i$ and $\boldsymbol{w}^T\boldsymbol{x}^- + b \leq -1 + \varepsilon_i$
$$\varepsilon_i \geq 0 \quad \forall i$$

Mapping functions: defining higher dimensions to achieve linear separation
$\sum_i \alpha_i y_i \varphi(\boldsymbol{x^i})^T\varphi(\boldsymbol{x^j}) + b$

Represent dot product as kernel: $K(\boldsymbol{x^i},\boldsymbol{x^j}) = \varphi(\boldsymbol{x^i})^T\varphi(\boldsymbol{x^j})$

MLE vs. MAP

Number of parameters:

Deriving parameter estimates: log-probability, set derivative to 0 (our Bayesian examples so far) or change parameter in direction of derivative (our logistic regression examples so far)

Gaussian

MLE
MAP