1. $$\tilde{x}_j = \frac{x_j - \mu_j}{\sigma_j}$$

2. the average squared distance from the data points to their assigned cluster centers

3. You'd use boosting. Bagging is meant to reduce variance, whereas boosting is meant to reduce the bias.

4. $$\mathcal{L}_{\mathrm{E}}(y, t) = \exp(-ty)$$

   Choose $\varepsilon = 1$. If the total loss is less than 1, then it must be less than 1 for each example. By inspection, the loss can only be less than 1 if the example is correctly classified.

5. Because the predictions are unbounded for Model 1, confident correct answers can be highly penalized. Model 2 avoids this effect.

6. $$Q^\pi(s, a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t \,\middle|\, S_0 = s, A_0 = a\right]$$
   $$Q^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(s'|s,a)}\left[Q^\pi(s', \pi(s'))\right]$$

7. If z is large enough, then y is numerically 1, and `np.log(1-y)` returns $-\infty$. Code to fix it:

```
def cross_entropy_loss(z, t):
    return t * np.logaddexp(0, -z) + (1-t) * np.logaddexp(0, z)
```

8. (a) TRUE. The hinge loss can only be zero if the example satisfies the margin constraint, and hence is classified correctly. This must be true for every example if the total hinge loss is zero.

   (b) The loss is minimized by $y = 0$, which can be achieved with $\mathbf{w} = 0$. Since this also minimizes the regularizer, it will choose $\mathbf{w} = 0$.

9.
$$\frac{1}{\beta}\sum_j |w_j| + \frac{1}{2\sigma^2}\sum_i (t^{(i)} - \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}))^2$$

The Laplace prior will encourage sparsity in the weights, since it corresponds to the L_1 norm.

10.
$$\begin{aligned}\overline{z_i} &= \phi'(z_i)\overline{y_i}\\ \overline{h_j} &= w_{ij}\overline{z_i}\\ \overline{w_{ij}} &= h_j\overline{z_i}\end{aligned}$$

11. E.g., $a = 75$, $b = 25$

12. Pick the first k columns of Q. For the derivation, see Lecture 12.

Sketch for other part

- $\text{Cov}(z) = U^T \text{Cov}(x) U$
- Use the spectral decomposition of Σ
- Use $U^T Q = \begin{pmatrix} I & 0 \end{pmatrix}$

13. (a)

$$\begin{aligned}\log p(\mathcal{D}_{\text{complete}}) &= \sum_{i=1}^{N} p(x^{(i)}, z^{(i)})\\ &= \sum_{i=1}^{N} (1 - z^{(i)})\left[\log(1-\theta) + \log \mathcal{N}(x^{(i)}; \mu, \sigma_0)\right] + z^{(i)}\left[\log\theta + \log\mathcal{N}(x^{(i)}; \mu, \sigma_1)\right]\end{aligned}$$

(b)

$$r^{(i)} = \frac{\theta\mathcal{N}(x^{(i)}; \mu, \sigma_1)}{(1-\theta)\mathcal{N}(x^{(i)}; \mu, \sigma_0) + \theta\mathcal{N}(x^{(i)}; \mu, \sigma_1)}$$

(c)

$$\sum_{i=1}^{N} (1 - r^{(i)})\left[\log(1-\theta) + \log \mathcal{N}(x^{(i)}; \mu, \sigma_0)\right] + r^{(i)}\left[\log(\theta) + \log\mathcal{N}(x^{(i)}; \mu, \sigma_1)\right]$$

(d) Considering only the parts of the objective which contain μ, we must maximize:

$$\sum_{i=1}^{N}(1-r^{(i)})\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma_0^2}\right)+r^{(i)}\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma_1^2}\right)$$

Differentiating with respect to μ and setting to zero:

$$0=\sum_{i=1}^{N}(1-r^{(i)})\left(\frac{x^{(i)}-\mu}{\sigma_0^2}\right)+r^{(i)}\left(\frac{x^{(i)}-\mu}{\sigma_1^2}\right)$$

After some rearranging, we get:

$$\mu=\frac{\frac{1}{\sigma_0^2}\sum_{i=1}^{N}(1-r^{(i)})x^{(i)}+\frac{1}{\sigma_1^2}\sum_{i=1}^{N}r^{(i)}x^{(i)}}{\frac{1}{\sigma_0^2}\sum_{i=1}^{N}(1-r^{(i)})+\frac{1}{\sigma_1^2}\sum_{i=1}^{N}r^{(i)}}$$

(e) Considering only the parts of the objective which contain σ_1, we must maximize:

$$\sum_{i=1}^{N}r^{(i)}\left[-\log(\sigma_1)+\left(-\frac{(x^{(i)}-\mu)^2}{2\sigma_1^2}\right)\right]$$

Differentiating with respect to σ_1 and setting to 0:

$$0=\sum_{i=1}^{N}r^{(i)}\left[-\frac{1}{\sigma_1}+\left(\frac{(x^{(i)}-\mu)^2}{\sigma_1^3}\right)\right]$$

After some rearranging, we get:

$$\sigma_1^2=\frac{\sum_{i=1}^{N}r^{(i)}(x^{(i)}-\mu)^2}{\sum_{i=1}^{N}r^{(i)}}$$