STAT C161/261 Equation Sheet

Likelihood, log likelihood, maximum likelihood estimation. Likelihood of full sample $\mathcal{X} = \{x^t\}_{t=1}^N$ when each sample is drawn from $p(x|\theta)$ distribution: $\ell(\theta|\mathcal{X}) \equiv p(\mathcal{X}|\theta) = \prod_{t=1}^N p(x^t|\theta)$. Log likelihood: $\mathcal{L}(\theta|\mathcal{X}) \equiv \log \ell(\theta|\mathcal{X}) = \sum_{t=1}^N \log p(x^t|\theta)$. Maximum likelihood (ML) estimate of θ from $\mathcal{X}$: θ that maximizes $\ell(\theta|\mathcal{X})$ or $\mathcal{L}(\theta|\mathcal{X})$.

Bias and variance. Estimator $d(\mathcal{X})$ has mean square error $r(d, m\theta) = E[(d(\mathcal{X}) - \theta)^2]$ and bias $b_\theta(d) = E[d(\mathcal{X}] - \theta$. Decomposition of MSE: $r(d, \theta) = \text{Var}(d) + (b_\theta(d))^2$.

Bayesian estimation. Given a prior density $p(\theta)$, the posterior density is $p(\theta|\mathcal{X}) = p(\mathcal{X}|\theta)p(\theta)/p(\mathcal{X})$. Maximum a posteriori (MAP) estimate: θ that maximizes $p(\theta|\mathcal{X})$. Bayes' estimator is $E[\theta|\mathcal{X}]$.

Multivariate normal distribution. $p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^\mathsf{T}\Sigma^{-1}(\mathbf{x}-\mu)\right]$.

Regression. A multivariate linear model is

$$\mathbf{r}^t = g(\mathbf{x}^t|w_0, w_1, \ldots, w_d) + \epsilon = w_0 + w_1 x_1^t + w_2 x_2^t + \cdots + w_d x_d^t + \epsilon.$$

After writing this as $\mathbf{r} = \mathbf{X}\mathbf{w} + \epsilon$, the least-squares solution is $\mathbf{w} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{r}$. If ϵ is i.i.d. with zero mean and variance σ^2, the error variance in $\mathbf{w}$ is $\text{Var}(\mathbf{w}) = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$.

Principal component analysis. $\Sigma = \text{Cov}(\mathbf{x})$. Seek $z_1 = \mathbf{w}_1^\mathsf{T}\mathbf{x}$ to maximize $\text{Var}(z_1)$ under constraint $\mathbf{w}_1^\mathsf{T}\mathbf{w}_1 = 1$. Solution is eigenvector associated with largest eigenvalue of Σ. Recurse. The proportion of variance explained by k principal components in a d-dimensional space: $(\lambda_1 + \lambda_2 + \cdots + \lambda_k)/(\lambda_1 + \lambda_2 + \cdots + \lambda_d)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are the eigenvalues of Σ.

Linear discriminant analysis (2 class). Samples of class i have mean $\mathbf{m}_i$, mean after projection $m_i = \mathbf{w}^\mathsf{T}\mathbf{m}_i$, and scatter after projection $s_i^2 = \sum(\mathbf{w}^\mathsf{T}\mathbf{x}^t - m_i)^2$ (where sum is over t such that $\mathbf{x}^t$ is in class i). Want to maximize Fisher's linear discriminant $J(\mathbf{w}) = (m_1 - m_2)^2/(s_1^2 + s_2^2)$. Solution is $\mathbf{w} = cS_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$, where $S_W = S_1 + S_2$ with S_i the within-class scatter matrix for class i: $S_i = \sum(\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^\mathsf{T}$ (where sum is over t such that $\mathbf{x}^t$ is in class i).

k-means clustering. Update of center of cluster i: $\mathbf{m}_i = \left(\sum_t b_i^t \mathbf{x}^t\right) / \left(\sum_t b_i^t\right)$ where b_i^t is the indicator of $\mathbf{x}^t$ being in cluster i. Update of cluster assignments: assign $\mathbf{x}^t$ to the closest cluster center.

Kernel density estimate. $\hat{p}(x) = \frac{1}{Nh}\sum_{t=1}^N K\left(\frac{x - x^t}{h}\right)$.

k-nearest neighbor classifier. Assign input to class having the most examples among the k nearest neighbors of the input.

Perceptron, training. Weights $\mathbf{w} = [w_0, w_1, \ldots, w_d]^\mathsf{T}$ and augmented input $\mathbf{x} = [1, x_1, \ldots, x_d]^\mathsf{T}$. Nonlinearity (but not always) sigmoid, giving $y = \text{sigmoid}(\mathbf{w}^\mathsf{T}\mathbf{x}) = \frac{1}{1+\exp[-\mathbf{w}^\mathsf{T}\mathbf{x}]}$. With sigmoid and second-layer weights denoted $\mathbf{v}$, stochastic gradient descent learning (backpropagation):

$$\begin{aligned}
\Delta v_h &= \eta \sum_t (r^t - y^t) z_h^t \\
\Delta w_{hj} &= \eta \sum_t (r^t - y^t) v_h z_h^t (1 - z_h^t) x_j^t
\end{aligned}$$