CSC 411
Machine Learning & Data Mining
Solutions

1 Locally reweighted regression

Given $\{(\mathbf{x}^{(1)}, y^{(1)}), .., (\mathbf{x}^{(N)}, y^{(N)})\}$ and positive weights $a^{(1)}, ..., a^{(N)}$ show that the solution to the *weighted* least square problem

$$\mathbf{w}^* = \arg\min \frac{1}{2} \sum_{i=1}^{N} a^{(i)} (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 + \frac{\lambda}{2} ||\mathbf{w}||^2 \tag{1}$$

is given by the formula

$$\mathbf{w}^* = \left(\mathbf{X}^T \mathbf{A} \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^T \mathbf{A} \mathbf{y} \tag{2}$$

where $\mathbf{X}$ is the design matrix (defined in class) and $\mathbf{A}$ is a diagonal matrix where $\mathbf{A}_{ii} = a^{(i)}$

1.1 Solution

Define the vector $\mathbf{r} = \mathbf{y} - \mathbf{X}\mathbf{w}$ then the first term in the loss can be written as $\frac{1}{2} \sum_{j=1}^{N} r_j^2 a^{(j)}$. If we look at $A\mathbf{r}$ we see that $[A\mathbf{r}]_j = a^{(j)} r_j$, so the inner product $\langle \mathbf{r}, A\mathbf{r} \rangle = \mathbf{r}^T A \mathbf{r} = \sum_j r_j \cdot a^{(j)} r_j = \sum_{j=1}^{N} r_j^2 a^{(j)}$. This means we can rewrite the loss as

$$\begin{aligned} L(\mathbf{w}) &= \frac{1}{2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T \mathbf{A} (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \frac{1}{2} ||\mathbf{w}||^2 \\ &= \frac{1}{2} \left(\mathbf{y}^T \mathbf{A} \mathbf{y} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{A} \mathbf{y} + \mathbf{w}^T \mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{w} + \lambda \mathbf{w}^T \mathbf{w} \right) \end{aligned}$$

using the same derivatives formulas we used in class, $\nabla_{\mathbf{w}} \mathbf{w}^T \mathbf{w} = 2\mathbf{w}$, $\nabla_{\mathbf{w}} \mathbf{w}^T A \mathbf{w} = 2A\mathbf{w}$ (holds for symmetric A) and $\nabla_{\mathbf{w}} \mathbf{w}^T \mathbf{x} = \mathbf{x}$ we get that

$$\nabla_{\mathbf{w}} L(\mathbf{w}) = -\mathbf{X}^T \mathbf{A} \mathbf{y} + \mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{w} + \lambda \mathbf{w}$$

Setting it to zero at the optimal $\mathbf{w}^*$ we get that

$$\mathbf{X}^T \mathbf{A} \mathbf{X} \mathbf{w}^* + \lambda \mathbf{w}^* = (\mathbf{X}^T \mathbf{A} \mathbf{X} + \lambda \mathbf{I}) \mathbf{w}^* = \mathbf{X}^T A \mathbf{y}$$

Multiplying both sides by $(\mathbf{X}^T \mathbf{A} \mathbf{X} + \lambda \mathbf{I})^{-1}$ (notice that it is positive-definite and therefore invertible) we get

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

2 Mini-batch SGD Gradient Estimator

Consider a dataset $\mathcal{D}$ of size n consisting of $(\mathbf{x}, y)$ pairs. Consider also a model $\mathcal{M}$ with parameters θ to be optimized with respect to a loss function $L(\mathbf{x}, y, \theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{x}^{(i)}, y^{(i)}, \theta)$.

We will aim to optimize L using mini-batches drawn randomly from $\mathcal{D}$ of size m. The indices of these points are contained in the set $\mathcal{I} = \{i_1, \ldots, i_m\}$, where each index is distinct and drawn uniformly without replacement from $\{1, \ldots, n\}$. We define the loss function for a single mini-batch as,

$$L_{\mathcal{I}}(\mathbf{x}, y, \theta) = \frac{1}{m} \sum_{i \in \mathcal{I}} \ell(\mathbf{x}^{(i)}, y^{(i)}, \theta) \tag{3}$$

1. Given a set $\{a_1, \ldots, a_n\}$ and random mini-batches $\mathcal{I}$ of size m, show that

$$\mathbb{E}_{\mathcal{I}} \left[\frac{1}{m} \sum_{i \in \mathcal{I}} a_i \right] = \frac{1}{n} \sum_{i=1}^{n} a_i$$

2.1 Solution

We can write,

$$\begin{aligned}
\mathbb{E}_{\mathcal{I}} \left[\frac{1}{m} \sum_{i \in \mathcal{I}} a_i \right] &= \mathbb{E}_{\mathcal{I}} \left[\frac{1}{m} \sum_{i=1}^{n} a_i \mathbb{1}[i \in \mathcal{I}] \right] \\
&= \frac{1}{m} \sum_{i=1}^{n} a_i \mathbb{P}(i \in \mathcal{I}) \\
&= \frac{1}{m} \frac{m}{n} \sum_{i=1}^{n} a_i = \frac{1}{n} \sum_{i=1}^{n} a_i
\end{aligned}$$

Noting that the probability of sampling a_i without replacement is $\frac{m}{n}$.

2. Show that $\mathbb{E}_{\mathcal{I}} \left[\nabla L_{\mathcal{I}}(\mathbf{x}, y, \theta) \right] = \nabla L(\mathbf{x}, y, \theta)$

2.2 Solution

Apply the above with $a_i = \nabla \ell(\mathbf{x}^{(i)}, y^{(i)}, \theta)$

3. Write, in a sentence, the importance of this result.

2.3 Solution

This tells us that SGD produces an unbiased estimate of the true gradient.

3 Class-Conditional Gaussians

In this question, you will derive the maximum likelihood estimates for class-conditional Gaussians with independent features (diagonal covariance matrices), i.e. Gaussian Naive Bayes, with shared variances. Start with the following generative model for a discrete class label $y \in (1, 2, ..., k)$ and a real valued vector of d features $\mathbf{x} = (x_1, x_2, ..., x_d)$:

$$p(y = k) = \alpha_k \tag{4}$$

$$p(\mathbf{x}|y = k, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \left(\prod_{i=1}^{D} 2\pi\sigma_i^2\right)^{-1/2} \exp\left\{-\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i - \mu_{ki})^2\right\} \tag{5}$$

where α_k is the prior on class k, σ_i^2 are the shared variances for each feature (in all classes), and μ_{ki} is the mean of the feature i conditioned on class k. We write $\boldsymbol{\alpha}$ to represent the vector with elements α_k and similarly $\boldsymbol{\sigma}$ is the vector of variances. The matrix of class means is written $\boldsymbol{\mu}$ where the kth row of $\boldsymbol{\mu}$ is the mean for class k.

1. Use Bayes' rule to derive an expression for $p(y = k|x, \boldsymbol{\mu}, \boldsymbol{\sigma})$.

3.1 Solution

$$p(y = k|\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{p(\mathbf{x}|y = k, \boldsymbol{\mu}, \boldsymbol{\sigma})p(y = k)}{\sum_k p(y = k)p(\mathbf{x}|y = k, \boldsymbol{\mu}, \boldsymbol{\sigma})} \tag{6}$$

$$= \frac{\left(\prod_{i=1}^{D} 2\pi\sigma_i^2\right)^{-1/2} \exp\left\{-\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i - \mu_{ki})^2\right\} \alpha_k}{\sum_k \left(\prod_{i=1}^{D} 2\pi\sigma_i^2\right)^{-1/2} \exp\left\{-\sum_{i=1}^{D} \frac{1}{2\sigma_i^2}(x_i - \mu_{ki})^2\right\} \alpha_k} \tag{7}$$

2. Write down an expression for the negative likelihood function (NLL)

$$\ell(\boldsymbol{\theta}; D) = -\log p(y^{(1)}, \mathbf{x}^{(1)}, y^{(2)}, \mathbf{x}^{(2)}, \cdots, y^{(N)}, \mathbf{x}^{(N)}|\boldsymbol{\theta}) \tag{8}$$

of a particular dataset $D = \{(y^{(1)}, \mathbf{x}^{(1)}), (y^{(2)}, \mathbf{x}^{(2)}), \cdots, (y^{(N)}, \mathbf{x}^{(N)})\}$ with parameters $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\sigma}\}$. (Assume that the data are iid.)

3.2 Solution

We write,

$$\log p(y^{(1)}, \mathbf{x}^{(1)}, \ldots, y^{(N)}, \mathbf{x}^{(N)}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p(\mathbf{x}^{(i)}|y^{(i)}, \boldsymbol{\theta}) + \log p(y^{(i)}|\boldsymbol{\theta}) \tag{9}$$

Substituting terms given in question yields result.

3. Take partial derivatives of the likelihood with respect to each of the parameters μ_{ki} and with respect to the shared variances σ_i^2.

3.3 Solution

Final form of derivatives as follows:

$$\frac{\partial(\cdots)}{\partial \boldsymbol{\mu}_{kj}} = -\sum_{i=1}^{N} \mathbb{1}[y^{(i)} = k](x_{ij} - \mu_{kj})\frac{1}{\sigma_j^2} \tag{10}$$

$$\frac{\partial(\cdots)}{\partial \boldsymbol{\sigma}_j^2} = \frac{-N}{2\sigma_j^2} + \sum_{i=1}^{N}(x_{ij} - \mu_{kj})^2 \frac{1}{2\sigma_j^4} \tag{11}$$

4. Find the maximum likelihood estimates for $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$.

3.4 Solution

Final solution (vectorized) is as follows:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[y^{(i)} = k]\mathbf{x}^{(i)} \tag{12}$$

$$\hat{\boldsymbol{\sigma}} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}})^2 \tag{13}$$

(Square taken elementwise in equation 13)

4 Kernels

In this question you will prove some properties of kernel functions. The two main ways to show a function $k(\mathbf{x}, \mathbf{y})$ is a kernel function is to find an embedding $\phi(x)$ such that $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$ or to show the for all $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ the gram matrix $K_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ is positive semi-definite (i.e. symmetric and no negative eigenvalues).

4.1 Positive semidefinite and quadratic form

1. Prove that a symmetric matrix $K \in \mathbb{R}^{d \times d}$ is positive semidefinite iff for all vectors $\mathbf{x} \in \mathbb{R}^d$ we have $\mathbf{x}^T K \mathbf{x} \geq 0$.

4.2 Solution

Proving $\Rightarrow$: If K is PSD then there exists a orthonoraml basic of eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_d$ with non-negative eigenvalues $\lambda_1, ..., \lambda_d$. For all vector $\mathbf{x}$ we can write it using these basis elements $\mathbf{x} = \sum_{i=1}^{d} a_i \mathbf{v}_i$. We now get

$$\mathbf{x}^T K \mathbf{x} = \left(\sum_i a_i \mathbf{v}_i\right)^T K \left(\sum_j a_j \mathbf{v}_j\right) = \sum_{i,j} a_i a_j \mathbf{v}_i^T K \mathbf{v}_j = \sum_{i,j} a_i a_j \mathbf{v}_i^T \lambda_j \mathbf{v}_j =$$
$$\sum_{i,j} a_i a_j \lambda_j \delta(i,j) = \sum_i a_i^2 \lambda_i \geq 0$$

Proving $\Leftarrow$: If the quadratic form is non-negative and $\mathbf{v}$ is an eigenvector with eigenvalue λ then

$$0 \leq \mathbf{v}^T K \mathbf{v} = \mathbf{v}^T \lambda \mathbf{v} = \lambda ||\mathbf{v}||^2 \Rightarrow \lambda \geq 0$$

4.3 Kernel properties

Prove the following properties:

1. The function $k(\mathbf{x}, \mathbf{y}) = \alpha$ is a kernel for $\alpha > 0$.
2. $k(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) \cdot f(\mathbf{y})$ is a kernel for all $f : \mathbb{R}^d \to \mathbb{R}$.
3. If $k_1(\mathbf{x}, \mathbf{y})$ and $k_2(\mathbf{x}, \mathbf{y})$ are kernels then $k(\mathbf{x}, \mathbf{y}) = a \cdot k_1(\mathbf{x}, \mathbf{y}) + b \cdot k_2(\mathbf{x}, \mathbf{y})$ for $a, b > 0$ is a kernel.
4. If $k_1(\mathbf{x}, \mathbf{y})$is a kernel then $k(\mathbf{x}, \mathbf{y}) = \frac{k_1(\mathbf{x}, \mathbf{y})}{\sqrt{k_1(\mathbf{x}, \mathbf{x})}\sqrt{k_1(\mathbf{y}, \mathbf{y})}}$ is a kernel (hint: use the features ϕ such that $k_1(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$).

4.4 Solution

1. $k(\mathbf{x}, \mathbf{y}) = \alpha$ corresponds to the feature mapping $\phi(\mathbf{x}) = \sqrt{\alpha}$: $\langle \phi(x), \phi(y) \rangle = \langle \sqrt{\alpha}, \sqrt{\alpha} \rangle = \alpha = k(\mathbf{x}, \mathbf{y})$. You can also show that $\mathbf{x}^T K \mathbf{x} = \alpha \sum_{ij} \mathbf{x}_i \mathbf{x}_j = \alpha ||\mathbf{x}||^2 \geq 0$.
2. $k(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) \cdot f(\mathbf{y})$ corresponds to the feature mapping $\phi(\mathbf{x}) = f(x) \in \mathbb{R}$.
3. If $k_1(\mathbf{x}, \mathbf{y})$ and $k_2(\mathbf{x}, \mathbf{y})$ are kernels then $k(\mathbf{x}, \mathbf{y}) = a \cdot k_1(\mathbf{x}, \mathbf{y}) + b \cdot k_2(\mathbf{x}, \mathbf{y})$ for $a, b > 0$ is a kernel - We have $K = aK_1 + bK_2$ so $\mathbf{x}^T K \mathbf{x} = \mathbf{x}^T (aK_1 + bK_2) \mathbf{x} = a \mathbf{x}^T K_1 \mathbf{x} + b \mathbf{x}^T K_2 \mathbf{x} \geq 0$.
4. $k_1(\mathbf{x}, \mathbf{y})$ is a kernel so there is some ϕ such that $k_1(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$. Define a new feature $\psi(\mathbf{x}) = \frac{\phi(\mathbf{x})}{||\phi(\mathbf{x})||} = \frac{\phi(\mathbf{x})}{\sqrt{k_1(\mathbf{x}, \mathbf{x})}}$ then $\langle \psi(\mathbf{x}), \psi(\mathbf{y}) \rangle = \frac{\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle}{\sqrt{k_1(\mathbf{x}, \mathbf{x})}\sqrt{k_1(\mathbf{y}, \mathbf{y})}} = \frac{k_1(\mathbf{x}, \mathbf{y})}{\sqrt{k_1(\mathbf{x}, \mathbf{x})}\sqrt{k_1(\mathbf{y}, \mathbf{y})}} = k(\mathbf{x}, \mathbf{y})$.